Energy Delay tradeoff in cooperative communication

Deepanshu Vasal
under supervision of
Prof. Achilleas Anastasopoulos
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor

Abstract

In a cooperative communication scenario, where a transmitter can transmit a packet directly to the receiver or indirectly through a relay, there is an inherent trade-off between energy and delay. While it may consume more energy to transmit a packet directly to the receiver than transmitting through a relay, the transmission through relay incurs more delay.

We pose this problem as an infinite horizon stochastic control problem. There are two possible cases: the centralized case where there exists a centralized controller which views the queue lengths of both the nodes, and the decentralized case where the information about the queue size of a node is not available to the other node. We find the optimum centralized control policy if the relay node does not have its own traffic, and show that, under certain conditions, it can be implemented in a decentralized fashion. For the decentralized case we consider traffic at both transmitter and the relay node and we prove a structural result that the optimum policy is the solution of a dynamic programming equation and the optimization is done over a fixed state space i.e., a state space that does not increase with time.

I. Introduction

In a wireless channel, successful communication between any two nodes is influenced by the channel statistics, transmission energy, energy path loss and interference by other users at the receiver, among other factors. With increasing number of wireless networking devices using real time applications, the delay is an important parameter for QoS (quality of service) of the communication, whereas due to battery constraints, the transmission energy is costly.

In a wireless network, the energy required to transmit a packet successfully to a receiver could be large due to large distance between the two nodes or bad channel gain, but presence of other nodes in the network could provide alternate route with possibly less energy costs. But since this requires successful transmission from the transmitter to the relay node and then from relay node to the receiver node, clearly the delay is more. Thus there is a tradeoff between the energy cost for successfully routing a packet and the delay cost.

In this work, we first consider a relay channel with a transmitter, relay and a receiver node with incoming traffic at the transmitter node only. There are fixed energy costs for any successful transmission from transmitter to relay, relay to receiver, or transmitter to the receiver, and there is a delay cost for each packet in the queue of either transmitter or the receiver. And the energy and delay costs are common knowledge among the nodes. There could be two possible cases regarding the information structure of the problem, centralized and decentralized. In the centralized case, the queue lengths of both, transmitter and relay node are common knowledge, whereas in the decentralized case, the queue length of any node is its private knowledge. The objective is to find the optimal strategy to be implemented by the transmitter and the relay node in the centralized and decentralized fashion that minimizes the total cost of energy and delay.

The remainder of this work is structured as follows. In section II, we present the model . In section III, we consider the case with traffic only at the transmitter node. We define the centralized control problem and find the optimum strategy for the centralized case and also show that it can be implemented in a decentralized case as well. In section IV we consider the decentralized scenario, under the assumption that the relay is also having its own traffic. We formulate the problem as an instance of decentralized control with delayed sharing pattern [4] and prove structural result that the optimum policy is the solution of a dynamic programming equation where the optimization is done over a fixed state space as opposed to an ever-increasing state-space in general.

II. Model

Our model consists of a transmitter node (node 1), a relay node (node 2) and a receiver node (node 3). The time is discretized into slots and we assume Bernoulli packet arrival processes

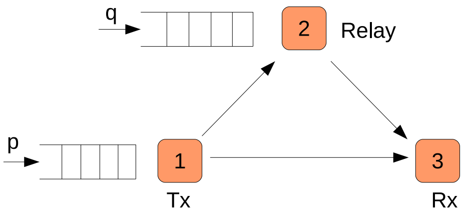

Fig. 1. A simple relay channel

p_t, q_t at node 1 and 2 respectively and the probability of arrival of a packet in any slot is p and q for node 1 and 2. Both node 1 and 2 have queues of infinite size. The transmitter has to send the arrived packets to the receiver and it has a choice to either transmit directly to the receiver or transmit it through the relay or not transmit at all. We denote by x_t^1 and x_t^2, the number of packets at time t in the queues of node 1 and node 2, respectively. Node 1 and node 2 take action u_t^1,u_t^2, respectively, as a function of all the information gathered till time t. The possible actions for node 1 are $\{E_{13}$:transmit to node 3, E_{12}: transmit to node 2, 0: wait (don't transmit) $\}$ and possible actions for node 2 are $\{E_{23}$: transmit to node 3, 0: wait $\}$. At the end of time slot t, node 1 and node 2 receive a noiseless feedback w_t from the receiver stating if the slot had successful transmission from node 1 (1) or node 2 (2), was idle (0), or had a collision (e). Thus each node at time t can determine (u_{t-1}^1, u_{t-1}^2) from its transmission at $t-1$ i.e. u_{t-1}^k and the feedback w_t and thus it is a delayed sharing of information with delay 1 [4].

The energy cost of transmission from node 1 to node 3 is E_{13}, node 1 to node 2 is E_{12} and that for node 2 to node 3 is E_{23} and simultaneous transmissions from both node 1 and node 2 lead to unsuccessful reception (collision), without any additional cost. To simplify notation, we consider the same symbols for actions as for the corresponding energy costs and reference is clear from context. We also assume a delay cost which is equal to the total number of packets waiting in the queues of node 1 and 2, thus cost of one unit per epoch for each packet in either queue. All costs are additive and costs for future slots (or epochs) are discounted by discount factor $\lambda, (0 < \lambda < 1)$. We describe $(E_{13}, E_{23}, E_{12}, \lambda, p, q)$ as the basic parameters of the system.

A. *Summary*

For all $k \in \{1,2\}, t \in \{1,2,...\}$

1) Queue lengths

   x_t^k : Queue length of node k at time t ; $x_t^k \in \{0,1,2,...\}$

2) Actions

   u_t^k : Action by node k at time t ; $u_t^1 \in \mathcal{U}^1 := \{0, E_{12}, E_{13}\}, u_t^2 \in \mathcal{U}^2 := \{0, E_{23}\}$

3) Feedback

   w_t : Feedback at time t (0:idle, 1:successful transmission from node 1, 2:successful transmission from node 2, e:collision) ; $w_t \in \{0,1,2,e\}$

4) Basic Random Variables

   $x_0^1, x_0^2, (p_t, q_t : t \in \{1,2...\})$ where p_t, q_t are Bernoulli arrival processes at each node's queue with parameters p and q respectively.

5) State Evolution

$$x_t^1 = p_t + x_{t-1}^1 - \mathbf{1}_{\{E_{12},E_{13}\}}(u_{t-1}^1)\mathbf{1}_{\{0\}}(u_{t-1}^2) \tag{1a}$$

$$x_t^2 = q_t + x_{t-1}^2 - \mathbf{1}_{\{E_{23}\}}(u_{t-1}^2)\mathbf{1}_{\{0\}}(u_{t-1}^1) + \mathbf{1}_{\{E_{12}\}}(u_{t-1}^1)\mathbf{1}_{\{0\}}(u_{t-1}^2) \tag{1b}$$

6) Instantaneous Cost

$$c_t(x_t^1, x_t^2, u_t^1, u_t^2) = x_t^1 + x_t^2 + u_t^1 + u_t^2 \tag{2}$$

7) Common Information at time t

   Since u_{t-1}^1 and w_t combined give u_{t-1}^2, we have common information as

$$u_{1:t-1}^{1:2} := u_1^1 u_1^2 u_2^1 u_2^2 \dots u_{t-1}^1 u_{t-1}^2$$

III. Centralized control with no traffic at relay

In this section, we are considering the centralized controlled system where queue length of both the transmitter and receiver nodes are known to the controller. We will also assume that there is no arrival process at the relay node, so $q = 0$. We will prove that the centralized policy can be implemented in decentralized way if the initial state of the system is $(0,0)$ i.e. $(x_1^1, x_1^2) = (0,0)$.

At time t, the common knowledge of nodes 1 and 2 (or centralized controller) is

$$\begin{aligned} u_{1:t-1}^{1:2} &:= u_1^1 u_1^2 u_2^1 u_2^2 \ldots u_{t-1}^1 u_{t-1}^2 \\ x_{1:t}^{1:2} &:= x_1^1 x_1^2 x_2^1 x_2^2 \ldots x_t^1 x_t^2 \end{aligned} \tag{3}$$

Thus the control action at time t , $u_t := (u_t^1, u_t^2) \in \mathcal{U}^1 \times \mathcal{U}^2$, in general, can be a function of all the information available till that time

$$\begin{aligned} u_t &= \hat{g}_t(x_{1:t}^{1:2}, u_{1:t-1})^1 \\ &= g_t(x_{1:t}^{1:2}) \end{aligned} \tag{4}$$

thus any policy **g** $= g_1, g_2, g_3, \cdots$ induces a cost

$$J^{\mathbf{g}} = \mathbb{E}\{\sum_{t=1}^{\infty} \lambda^{t-1} c_t(X_t^1, X_t^2, U_t^1, U_t^2)\} \tag{5}$$

The objective is to minimize the total discounted cost of energy and delay incurred over infinite time horizon. We define the problem as follows

Problem 1. *Find the optimum centralized policy* $\boldsymbol{g}^*$ *that achieves the optimum cost,*

$$J^* := \min_{\boldsymbol{g}} J^{\boldsymbol{g}} \tag{6}$$

where $J^{\boldsymbol{g}}$ *is as defined in (5), control actions* $u_t^{1:2}$ *as in (4).*

Lemma 1. *The process* $\{X_t^{1:2}, t = 0, 1, \cdots\}$ *is a controlled Markov process with control* U_t *and instantaneous cost as given in (2) i.e.*

$$\mathbb{P}(x_{t+1}^{1:2} | x_{1:t}^{1:2}, u_{1:t}^{1:2}) = \mathbb{P}(x_{t+1}^{1:2} | x_t^{1:2}, u_t^{1:2}) \tag{7}$$

Proof: This is trivially true due to system evolution as given in (1) and the independence of the basic random variables $(X_1^1, X_1^2, P_1, P_2 \cdots,)$. ■

Thus by Markov Decision Theory [1] [2], there exists a stationary Markov policy of the form

[1] In the text we repeatedly use functions $\hat{g}$ and $\hat{c}$ to emphasize its arguments and same notation should not be interpreted as the same functional form

$u_t = g(x_t^{1:2})$ that achieves optimum cost J^* as given in (6). Moreover this optimal cost can be found as

$$J^* \;=\; \mathbb{E}\{V(X_1^1, X_1^2)\} \tag{8}$$

where the cost-to-go function $V(x,y)$ satisfies the following dynamic programming equation (9) and the actions u_1, u_2 that achieves the minima in (9) for each state x, y form stationary optimal Markov policy.

$$V(x,y) = \min_{u_1,u_2} \begin{cases} (0,0) : x + y + \lambda p V(x+1,y) + \lambda(1-p)V(x,y) & (x \geq 0, y \geq 0) \\ (E_{13},0) : x + y + E_{13} + \lambda p V(x,y) + \lambda(1-p)V(x-1,y) & (x \geq 1, y \geq 0) \\ (E_{12},0) : x + y + E_{12} + \lambda p V(x,y+1) + \lambda(1-p)V(x-1,y+1) & (x \geq 1, y \geq 0) \\ (0,E_{23}) : x + y + E_{23} + \lambda p V(x+1,y-1) + \lambda(1-p)V(x,y-1) & (x \geq 0, y \geq 1) \\ (E_{13},E_{23}) : x + y + E_{13} + E_{23} + \lambda p V(x+1,y) + \lambda(1-p)V(x,y) & (x \geq 1, y \geq 1) \\ (E_{12},E_{23}) : x + y + E_{12} + E_{23} + \lambda p V(x+1,y) + \lambda(1-p)V(x,y) & (x \geq 1, y \geq 1) \end{cases} \tag{9}$$

A. *Solving the dynamic programming equation*

It can be easily seen that cost for actions (E_{12}, E_{23}) and (E_{13}, E_{23}) is always greater than that for $(0,0)$, thus it need not be considered in computing the minima. These actions lead to collision and the centralized control avoid that. Thus the cost-to-go function should satisfy

$$V(x,y) = \min_{u_1,u_2} \begin{cases} (0,0) : x + y + \lambda p V(x+1,y) + \lambda(1-p)V(x,y) & (x \geq 0, y \geq 0) \\ (E_{13},0) : x + y + E_{13} + \lambda p V(x,y) + \lambda(1-p)V(x-1,y) & (x \geq 1, y \geq 0) \\ (E_{12},0) : x + y + E_{12} + \lambda p V(x,y+1) + \lambda(1-p)V(x-1,y+1) & (x \geq 1, y \geq 0) \\ (0,E_{23}) : x + y + E_{23} + \lambda p V(x+1,y-1) + \lambda(1-p)V(x,y-1) & (x \geq 0, y \geq 1) \end{cases} \tag{10}$$

In general its difficult to solve such recursive equation to find $V(x,y)$. Moreover there could exist multiple solutions. As shown in [3, sec. 6.10] and appendix B, in appropriate Banach space, existence of a unique solution is guaranteed under certain conditions. We propose policies for different set of values of the basic parameters and prove their optimality by proving that they satisfy the dynamic programming equation (10). The uniqueness of the solution (in appropriate

Banach space) guarantees that the solution of the dynamic program is equal to the optimal solution.

1) Optimum policies under a restricted class of policies: First we restrict ourselves to the set of policies where whenever $x_t^2 > 0$ (i.e. queue of node 2 is non-empty), node 2 transmits its packets to node 3 and during this time node 1 waits. Also we assume that system starts at $x_t^1 = 0, x_t^2 = 0$ and as a consequence of the restriction on policies, in future $x_t^2 \in \{0, 1\}$. Thus for these restricted set of policies, the following equations have to be satisfied

$$V(x,1) = \quad x + 1 + E_{12} + \lambda p V(x+1,0) + \lambda(1-p)V(x,0) \tag{11}$$

$$V(x,0) = \quad \min_{u_1} \begin{cases} (0,0) : x + \lambda p V(x+1,0) + \lambda(1-p)V(x,0) \\ (E_{13},0) : x + E_{13} + \lambda p V(x,0) + \lambda(1-p)V(x-1,0) & (x>0) \\ (E_{12},0) : x + E_{12} + \lambda p V(x,1) + \lambda(1-p)V(x-1,1) & (x>0) \end{cases} \tag{12}$$

Substituting (11) into (12) we can restrict the state of the system to x, the queue size of node 1, which sufficiently describes the evolution of the system under restricted set of policies and thus the optimum cost-to-go function defined with slight abuse of notation as $V(x) := V(x,0)$, should satisfy the following equation

$$\begin{aligned} V(0) =& \lambda p V(1) + \lambda(1-p)V(0) \\ V(x) =& \min_{u_1} \begin{cases} (0) : x + \lambda p V(x+1) + \lambda(1-p)V(x) \\ (E_{13}) : x + E_{13} + \lambda p V(x) + \lambda(1-p)V(x-1) & (x>0) \\ (E_{12}) : (1+\lambda) + (E_{12} + \lambda E_{23} + \lambda p) + \lambda^2 p^2 V(x+1) \\ \qquad\qquad + \lambda^2(1-p)^2 V(x-1) + 2\lambda^2 p(1-p)V(x) & (x>0) \end{cases} \end{aligned}$$

Table I, II, III, IV shows the optimum policies for the case of restricted policies, the corresponding cost to go function and the set of parameters for which it is optimal. Appendix A contains the definitions of symbols used and Appendix C derives cost-to-go function for case I, II, III. We omit the proof of IV.

TABLE I
WHEN ITS OPTIMUM TO WAIT.

Policy	$x \geq 0 : (0, 0)$
Cost to go function	$V(x) = \frac{\lambda p}{(1-\lambda)^2} + sx$
Conditions of optimality	$E_{13} \geq \frac{\lambda}{1-\lambda}$ $E_{12} + \lambda E_{23} \geq \frac{\lambda^2}{1-\lambda}$

TABLE II
WHEN ITS OPTIMUM TO TRANSMIT DIRECTLY TO NODE 3

Policy	$x = 0 : (0, 0)$ $x > 0 : (E_{13}, 0)$
Cost to go function	$V(x) = c + sx + d\rho^x$
Conditions of optimality	$E_{13} \leq \frac{\lambda}{1-\lambda}$ $E_{13} \leq \frac{(1-\lambda p)}{(1-\lambda^2 p)}(E_{12} + \lambda E_{23}) + \frac{\lambda}{(1-\lambda^2 p)}$

TABLE III
WHEN ITS OPTIMUM TO TRANSMIT THROUGH RELAY NODE 2

Policy	$x = 0 : (0, 0)$ $x > 0 : (E_{12}, 0)$
Cost to go function	$V(x) = u + sx + r\xi^x$
Conditions of optimality	$E_{12} + \lambda E_{23} \leq \frac{\lambda^2}{1-\lambda}$ $\frac{E_{12}+\lambda E_{23}}{1+\lambda} + \frac{\lambda}{1-\lambda^2} \leq E_{13}$

TABLE IV
TRANSMIT THROUGH NODE 2 IF x (QUEUE SIZE OF NODE 1) IS STRICTLY LESS THAN x_{th} AND WE TRANSMIT DIRECTLY, IF x IS GREATER THAN EQUAL TO x_{th}

Policy	$x = 0 : (0, 0)$ $0 < x < x_{th} : (E_{12}, 0)$ $x_{th} \leq x : (E_{13}, 0)$
Cost to go function	$\forall x \leq x_{th} - 1 : \quad V(x) = r\xi^x + sx + u - l(a\alpha)^{x_{th}-x} w_1(x)$ $x = x_{th} - 1 : \; V(x_{th} - 1) = s(x_{th} - 1) + u + \frac{r\xi^{x_{th}-1} - (u-c)(1-\rho) w_1(x_{th}-1)(a\alpha)}{1-(\rho - a\beta) w_1(x_{th}-1) a\alpha}$ $\forall x \geq x_{th} : \quad V(x) = sx + c + \rho^{x - x_{th}+1}[V(x_{th} - 1) - s(x_{th} - 1) - c]$
Conditions of optimality	$(E_{13} - \frac{\lambda}{1-\lambda}) < \frac{b_{x_{th}+1}}{a_{x_{th}+1}}(E_{12} + \lambda E_{23} - \frac{\lambda^2}{1-\lambda})$ $(E_{13} - \frac{\lambda}{1-\lambda}) \geq \frac{b_{x_{th}}}{a_{x_th}}(E_{12} + \lambda E_{23} - \frac{\lambda^2}{1-\lambda})$

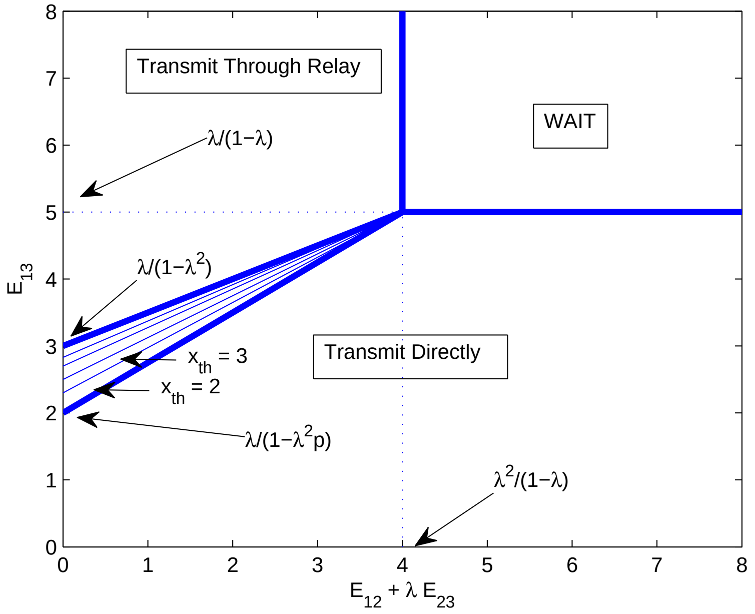

Fig. 2. Decision Regions in the space of basic parameters $E_{13}, E_{12} + \lambda E_{23}, \lambda, p$ for restricted set of policies. The triangular region represents policy described in Table IV with different x_{th}

Figure 2 shows the optimum policy for different decision regions in the space of basic parameters $E_{12}, E_{23}, E_{13}, \lambda, p$. These regions are also verified by the numerical analysis using the method of value iteration [3].

Figure 3 shows the decision regions in the limiting case as the discount factor $\lambda \to 1$.

2) Centralized policy for the general case: In the previous section we considered the restricted set of policies for which node 2 transmitted whenever it had a packet and thus $x_t^2 \in \{0, 1\}$. In this section, the policies for the general setting are proposed. The state of the system is (x,y) where ‘x’ and ‘y’ are the queue lengths of node 1 and node 2. The optimum policy has to satisfy equation (10). For the general case when there is no traffic at relay, Table V, VI, VII, VIII, IX, X show the optimum policies , the corresponding cost to go function and the set of parameters for which it is optimal.

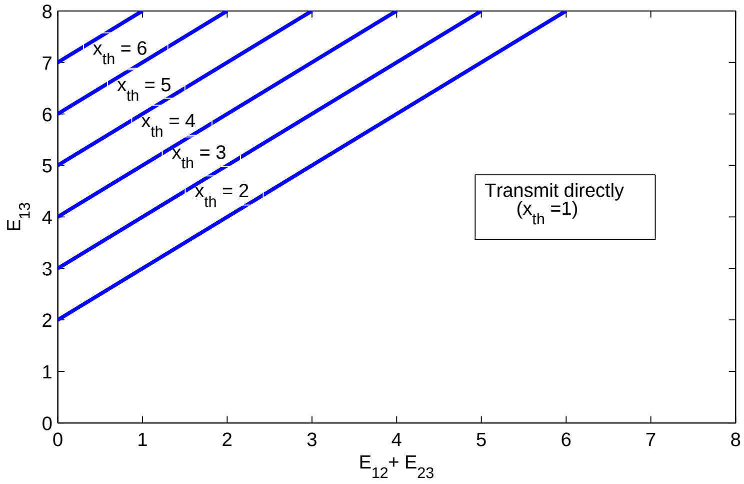

Fig. 3. Decision Regions in the space of basic parameters $E_{13}, E_{12} + \lambda E_{23}, \lambda, p$ for restricted set of policies as $\lambda \to 1$.

TABLE V
BOTH NODE 1 AND 2 WAIT

Policy	$x \geq 0, y \geq 0 : (0,0)$
Cost to go function	$V(x,y) = s(x+y) + \frac{\lambda p}{(1-\lambda)^2}$
Conditions of optimality	$E_{13} \geq \frac{\lambda}{1-\lambda}$ $E_{23} \geq \frac{\lambda}{1-\lambda}$

TABLE VI
NODE 1 HAS PRIORITY: FIRST NODE 1 TRANSMITS DIRECTLY TO 3 AND THEN NODE 2 TRANSMITS ITS PACKET TO 3

Policy	$x = 0, y = 0 : (0,0)$ $x = 0, y \geq 1 : (0, E_{23})$ $x \geq 1, y \geq 0 : (E_{13}, 0)$
Cost to go function	$V(x,y) = s(x + y + E_{13}) - \frac{\lambda(1-p)}{(1-\lambda)^2} + \rho^x \left[\frac{(1-\lambda p)(E_{23}-E_{13})}{1-\lambda} + \rho^y \left[\frac{(1-\lambda p)(\lambda - E_{23}(1-\lambda))}{(1-\lambda)^2} \right] \right]$
Conditions of optimality	$E_{13} \leq E_{23} \leq \frac{\lambda}{1-\lambda}$

TABLE VII
NODE 1 HAS PRIORITY: NODE 1 TRANSMITS ALL ITS PACKETS DIRECTLY TO 3 AND NODE 2 WAITS

Policy	$x = 0, y \geq 0 : (0,0)$ $x \geq 1, y \geq 0 : (E_{13}, 0)$
Cost to go function	$V(x,y) = s(x+y) + c + d\rho^x$
Conditions of optimality	$E_{13} \leq \frac{\lambda}{1-\lambda} \leq E_{23}$

TABLE VIII
NODE 2 HAS PRIORITY: NODE 1 WAITS WHILE NODE 2 TRANSMITS

Policy	$x \geq 0, y = 0 : (0,0)$ $x \geq 0, y \geq 1 : (0, E_{23})$
Cost to go function	$V(x,y) = s(x+y+E_{23}) - \frac{\lambda(1-p)}{(1-\lambda)^2} + \lambda^y \left[\frac{\lambda}{(1-\lambda)^2} - sE_{23} \right]$
Conditions of optimality	$E_{13} \geq \frac{\lambda}{1-\lambda}$ $\left(\frac{\lambda}{1-\lambda} - \frac{E_{12}}{\lambda} \right)^+ \leq E_{23} \leq \frac{\lambda}{1-\lambda}$

TABLE IX
NODE 2 HAS PRIORITY: FIRST NODE 2 TRANSMITS ALL ITS PACKETS TO NODE 3, THEN NODE 1 TRANSMITS TO 3

Policy	$x = 0, y = 0 : (0,0)$ $x \geq 0, y = 0 : (E_{13}, 0)$ $x \geq 0, y \geq 1 : (0, E_{23})$
Cost to go function	$V(x,y) = s(x+y+E_{23}) - \frac{\lambda(1-p)}{(1-\lambda)^2} + \lambda^y \left[\frac{E_{13}-E_{23}}{1-\lambda} + d\rho^x (p\rho + 1 - p)^y \right]$
Conditions of optimality	$E_{13} \leq \frac{(1-\lambda p)(E_{12}+\lambda E_{23})}{1-\lambda^2 p} + \frac{\lambda}{1-\lambda^2 p}$

TABLE X
NODE 2 HAS PRIORITY: NODE 2 TRANSMITS TO NODE 3 WHENEVER ITS QUEUE IS NONEMPTY AND NODE 1 TRANSMITS ITS PACKETS TO 2 IF NODE 2'S QUEUE IS EMPTY

Policy	$x = 0, y = 0 : (0,0)$ $x \geq 1, y = 0 : (E_{12}, 0)$ $x \geq 0, y \geq 1 : (0, E_{23})$
Cost to go function	$V(x,y) = s(x+y+E_{23}) - \frac{\lambda(1-p)}{(1-\lambda)^2} + \lambda^y \left[\frac{(E_{12}-E_{23})(1-\lambda)+\lambda}{(1-\lambda)^2(1+\lambda)} + r\xi^x (p\xi + 1 - p)^y \right]$
Conditions of optimality	$E_{13} \geq \frac{E_{12}+\lambda E_{23}}{1+\lambda} + \frac{\lambda}{1-\lambda^2}$ $\frac{\lambda^2}{1-\lambda} \geq E_{12} + \lambda E_{23}$

If the system starts from $(x_1^1, x_1^2) = (0, 0)$, then the general policy reduces to the restricted policies considered before and since packet arrival at node 2 is only through node 1, thus node 1 can keep a track of the queue size of node 2 and the optimum policy can be decentralized.

IV. Decentralized Control with incoming traffic at relay

In this section, we assume that the relay also has incoming traffic modelled as a Bernoulli arrival process with arrival probability, $q \in [0, 1]$. We consider the decentralized case where node 1 cannot observe the queue length of node 2 and vice versa. In this case either of the nodes cannot track the queue size of the other node and the problem becomes considerably more complex.

At time t, information available with node k is $(x_{1:t}^k, u_{1:t-1}^k, w_{1:t-1})$ which is equivalent to $(x_{1:t}^k, u_{1:t-1}^{1:2})$ and thus control actions can be defined as follows

$$\begin{aligned} u_t^1 &= \hat{g}_t^1(x_{1:t}^1, u_{1:t-1}^1, w_{1:t-1}) = g_t^1(x_{1:t}^1, u_{1:t-1}^{1:2}) \\ u_t^2 &= \hat{g}_t^2(x_{1:t}^2, u_{1:t-1}^2, w_{1:t-1}) = g_t^2(x_{1:t}^2, u_{1:t-1}^{1:2}) \end{aligned} \tag{13}$$

If $\mathbf{g}^k$ is any strategy of node k i.e. $\mathbf{g}^k = g_1^k, g_2^k, \cdots$ where $k \in \{1, 2\}$ then $\mathbf{g} = (\mathbf{g}^1, \mathbf{g}^2)$ is the combined strategy of both the nodes and the corresponding cost is given by $J^{\mathbf{g}}$

$$J^{\mathbf{g}} = \mathbb{E}\{\sum_{i=1}^{\infty} \lambda^{t-1} c_t(X_t^1, X_t^2, U_t^1, U_t^2)\} \tag{14}$$

Problem 2. *Find the optimum decentralized policy $\boldsymbol{g}^*$ that achieves the optimum cost,*

$$J^* := \min_{\boldsymbol{g}} J^{\boldsymbol{g}} \tag{15}$$

where $J^{\boldsymbol{g}}$ is as defined in (14), control actions $u_t^{1:2}$ as in (13).

Here we prove a structural result for the optimum decentralized policy and show that it can be found as a solution of a dynamic programming equation. First we prove that there exist optimum control actions that depend only on current state and entire control history i.e. $(x_t^k, u_{1:t-1}^{1:2})$ and further it depends only on the current state x_t^k and the posterior on $x_t^{1:2}$ conditioned on the control history $u_{1:t-1}^{1:2}$.

Lemma 2. *For any fixed strategy $\boldsymbol{g}$, random variables $X_{1:t}^1$ and $X_{1:t}^2$ are conditionally independent* *given the control history till time t, $U^{1:2}_{1:t-1}$ i.e.*

$$\mathbb{P}^{\boldsymbol{g}}(x^{1:2}_{1:t}|u^{1:2}_{1:t-1}) = \mathbb{P}^{\boldsymbol{g}^1}(x^1_{1:t}|u^{1:2}_{1:t-1})\mathbb{P}^{\boldsymbol{g}^2}(x^2_{1:t}|u^{1:2}_{1:t-1}) \tag{16}$$

Proof:

The causal decomposition of $\mathbb{P}^{\mathbf{g}}(x^{1:2}_{1:t}, u^{1:2}_{1:t-1})$ gives,

$$\begin{aligned}\mathbb{P}^{\mathbf{g}}(x^{1:2}_{1:t}, u^{1:2}_{1:t-1}) =& \mathbb{P}(x^1_1)\prod_{i=1}^{t-1}\Bigg(\mathbb{P}(x^1_{i+1}|x^1_i, u^{1:2}_i)\mathbb{P}^{\mathbf{g}^1}(u^1_i|x^1_{1:i}, u^{1:2}_{1:i-1})\Bigg) \\ & \mathbb{P}(x^2_1)\prod_{j=1}^{t-1}\Bigg(\mathbb{P}(x^2_{j+1}|x^2_j, u^{1:2}_j)\mathbb{P}^{\mathbf{g}^2}(u^2_j|x^2_{1:j}, u^{1:2}_{1:j-1})\Bigg)\end{aligned} \tag{17}$$

$$\begin{aligned}\mathbb{P}^{\mathbf{g}}(x^{1:2}_{1:t}|u^{1:2}_{1:t-1}) =& \frac{\mathbb{P}(x^1_1)\prod_{i=1}^{t-1}\Bigg(\mathbb{P}(x^1_{i+1}|x^1_i, u^{1:2}_i)\mathbb{P}^{\mathbf{g}^1}(u^1_i|x^1_{1:i}, u^{1:2}_{1:i-1})\Bigg)}{\sum_{x^1_{1:t}}\mathbb{P}(x^1_1)\prod_{i=1}^{t-1}\Bigg(\mathbb{P}(x^1_{i+1}|x^1_i, u^{1:2}_i)\mathbb{P}^{\mathbf{g}^1}(u^1_i|x^1_{1:i}, u^{1:2}_{1:i-1})\Bigg)} \\ & \frac{\mathbb{P}(x^2_1)\prod_{j=1}^{t-1}\Bigg(\mathbb{P}(x^2_{j+1}|x^2_j, u^{1:2}_j)\mathbb{P}^{\mathbf{g}^2}(u^2_j|x^2_{1:j}, u^{1:2}_{1:j-1})\Bigg)}{\sum_{x^2_{1:t}}\mathbb{P}(x^2_1)\prod_{j=1}^{t-1}\Bigg(\mathbb{P}(x^2_{j+1}|x^2_j, u^{1:2}_j)\mathbb{P}^{\mathbf{g}^2}(u^2_j|x^2_{1:j}, u^{1:2}_{1:j-1})\Bigg)} \qquad (18) \\ =& \mathbb{P}^{\mathbf{g}^1}(x^1_{1:t}|u^{1:2}_{1:t-1})\mathbb{P}^{\mathbf{g}^2}(x^2_{1:t}|u^{1:2}_{1:t-1}) \qquad (19)\end{aligned}$$

■

Lemma 3. *For given any fixed strategy of the node 2 i.e.* $\boldsymbol{g}^2$, $\{(X^1_t, U^{1:2}_{1:t-1}); t = 1, 2, \cdots\}$ *is a controlled Markov process with state ($X^1_t, U^{1:2}_{1:t-1}$) and control input U^1_t i.e.*

$$\mathbb{P}^{\boldsymbol{g}^2}(x^1_{t+1}, u^{1:2}_{1:t}|x^1_{1:t}, u^{1:2}_{1:t-1}, u^1_{1:t}) = \mathbb{P}^{\boldsymbol{g}^2}(x^1_{t+1}, u^{1:2}_{1:t}|x^1_t, u^{1:2}_{1:t-1}, u^1_t) \tag{20}$$

$$\begin{aligned}\mathbb{E}^{\boldsymbol{g}^2}\{c_t(x^1_t, x^2_t, u^1_t, u^2_t)|x^1_{1:t}, u^{1:2}_{1:t-1}, u^1_{1:t}\} &= \mathbb{E}^{\boldsymbol{g}^2}\{c_t(x^1_t, x^2_t, u^1_t, u^2_t)|x^1_t, u^{1:2}_{1:t-1}, u^1_t\} \qquad (21) \\ &= \hat{c}(x^1_t, u^{1:2}_{1:t-1}, u^1_t)\end{aligned}$$

Proof:

$$\begin{aligned}\mathbb{P}^{\mathbf{g}^2}(x^1_{t+1}, u^{1:2}_{1:t}|x^1_{1:t}, u^{1:2}_{1:t-1}, u^1_{1:t}) &= \mathbb{P}^{\mathbf{g}^2}(x^1_{t+1}|x^1_{1:t}, u^{1:2}_{1:t}).\mathbb{P}^{\mathbf{g}^2}(u^{1:2}_{1:t}|x^1_{1:t}, u^{1:2}_{1:t-1}, u^1_t) \\ &= \mathbb{P}^{\mathbf{g}^2}(x^1_{t+1}|x^1_{1:t}, u^{1:2}_{1:t}).\mathbb{P}^{\mathbf{g}^2}(u^2_t|x^1_{1:t}, u^{1:2}_{1:t-1}, u^1_t) \qquad (22)\end{aligned}$$

Since $x^1_{t+1} = f_t(x^1_t, p_{t+1}, u^{1:2}_t)$ where f_t is as defined in (1)

$$\begin{aligned}\mathbb{P}^{\mathbf{g}^2}(x^1_{t+1}|x^1_{1:t}, u^{1:2}_{1:t}) &= \sum_{p_{t+1}} \mathbb{P}^{\mathbf{g}^2}(x^1_{t+1}, p_{t+1}|x^1_{1:t}, u^{1:2}_{1:t}) \\ &= \sum_{p_{t+1}} \mathbb{P}(p_{t+1}) \mathbf{1}_{f_t(x^1_t, p_{t+1}, u^{1:2}_t)}\{x^1_{t+1}\} \\ &= \mathbb{P}(x^1_{t+1}|x^1_t, u^{1:2}_t) \qquad (23)\end{aligned}$$

also since U^2_t is a function of $X^2_{1:t}$, U^1_t is a function of $X^1_{1:t}$ and $X^1_{1:t}, X^2_{1:t}$ are conditionally independent given $U^{1:2}_{1:t-1}$ (Lemma 2), thus

$$\mathbb{P}^{\mathbf{g}^2}(u^2_t|x^1_{1:t}, u^{1:2}_{1:t-1}, u^1_t) = \mathbb{P}^{\mathbf{g}^2}(u^2_t|u^{1:2}_{1:t-1}) \qquad (24)$$

Thus from (22),(23) and (24),

$$\mathbb{P}^{\mathbf{g}^2}(x^1_{t+1}, u^{1:2}_{1:t}|x^1_{1:t}, u^{1:2}_{1:t-1}, u^1_{1:t}) = \mathbb{P}^{\mathbf{g}^2}(x^1_{t+1}, u^{1:2}_{1:t}|x^1_t, u^{1:2}_{1:t-1}, u^1_t) \qquad (25)$$

For the second part,

$$\begin{aligned}\mathbb{E}^{\mathbf{g}^2}\{c_t(x^1_t, x^2_t, u^1_t, u^2_t)|x^1_{1:t}, u^{1:2}_{1:t-1}, u^1_{1:t}\} &= \sum_{x^1_t, x^2_t, u^1_t, u^2_t} c_t(x^1_t, x^2_t, u^1_t, u^2_t)\mathbb{P}^{\mathbf{g}^2}(x^1_t, x^2_t, u^1_t, u^2_t|x^1_{1:t}, u^{1:2}_{1:t-1}, u^1_{1:t}) \qquad (26a) \\ &= \sum_{x^2_t, u^2_t} c_t(x^1_t, x^2_t, u^1_t, u^2_t)\mathbb{P}^{\mathbf{g}^2}(x^2_t, u^2_t|x^1_{1:t}, u^{1:2}_{1:t-1}, u^1_{1:t}) \qquad (26b) \\ &= \sum_{x^2_t, u^2_t} c_t(x^1_t, x^2_t, u^1_t, u^2_t)\mathbb{P}^{\mathbf{g}^2}(x^2_t, u^2_t|u^{1:2}_{1:t-1}) \qquad (26c) \\ &= \mathbb{E}^{\mathbf{g}^2}\{c_t(x^1_t, x^2_t, u^1_t, u^2_t)|x^1_t, u^{1:2}_{1:t-1}, u^1_t\} \qquad (26d) \\ &= \hat{c}(x^1_t, u^{1:2}_{1:t-1}, u^1_t) \qquad (26e)\end{aligned}$$

where (26c) is true since X^2_t, U^2_t are conditionally independent of $X^1_{1:t}, U^1_{1:t}$ given $U^{1:2}_{1:t-1}$ (Lemma 2). ■

As a consequence of the MDP structure of the problem, given a fixed strategy $\mathbf{g}^2$ of the node 2, the optimum control action by node 1 can be given as (for $k = 1$)

$$u^k_t = g^k_t(x^k_t, u^{1:2}_{1:t-1}) \qquad (27)$$

Since this is true for any fixed strategy of node 2, it is also true for the optimal strategy of the node 2. A similar result for node 2 is true, thus the above equation is valid for $k \in \{1,2\}$.

A. POMDP from the perspective of coordinator

In the decentralized case, each node can act as a controller and thus we have two linked stochastic control problems. We can view this problem from the perspective of a fictitious coordinator [4] who observes, at time t, the feedback w_t or equivalently $u_{t-1}^{1:2}$ but does not observe $x_t^k, k \in \{1,2\}$. Thus at time t, it has access to the information $u_{1:t-1}^{1:2}$ (due to perfect recall) and based upon this information, it generates partial functions $\gamma_t^{1:2}$ as its control output, where $\gamma_t^k : \mathbb{N} \rightarrow \mathcal{U}^k, k \in \{1,2\}$. And based upon these control outputs of the coordinator, node k, $k \in \{1,2\}$ compute its action by operating these partial functions on its private information i.e. x_t^k . If strategy of the coordinator is Ψ, then

$$\begin{aligned} (\gamma_t^1, \gamma_t^2) &= \Psi_t(u_{1:t-1}^{1:2}) & (28)\\ u_t^k &= \gamma_t^k(x_t^1) = \Psi_t^k(u_{1:t-1}^{1:2})(x_t^k) = g_t^k(u_{1:t-1}^{1:2}, x_t^k) & (29) \end{aligned}$$

Now we show that belief on $x_t^{1:2}$ given the observation and control history till time t which is $u_{1:t-1}^{1:2}, \gamma_{1:t-1}^{1:2}$, forms a sufficient state for the coordinator's problem. We define the random variable $\Pi_t \in \mathcal{P}(\mathbb{N}^2)$ as the posterior pmf of $X_t^{1:2}$ conditioned on $U_{1:t}^{1:2}, \Gamma_{1:t-1}^{1:2}$ i.e.

$$\Pi_t(x_t^{1:2}) = \mathbb{P}(X_t^{1:2} = x_t^{1:2} | U_{1:t-1}^{1:2}, \Gamma_{1:t-1}^{1:2}) \tag{30}$$

Lemma 4.

$$\pi_{t+1} = F(\pi_t, \gamma_t^{1:2}, u_t^{1:2}) \tag{31}$$

where F is a deterministic update function that does not depend upon the policy ***g***

Proof: See Appendix D ■

Proposition 1. *The process* $\{\Pi_t, t = 1, 2, ...T\}$ *is a controlled Markov Process with control* $\gamma_t^{1:2}$.

i.e.

$$\begin{aligned}\mathbb{P}(\pi_{t+1}|\pi_{1:t},\gamma_{1:t}^{1:2}) &= \mathbb{P}(\pi_{t+1}|\pi_t,\gamma_t^{1:2}) \quad (32)\\ \mathbb{E}(c(x_t^{1:2},u_t^{1:2})|\pi_{1:t},\gamma_{1:t}^{1:2}) &= \mathbb{E}(c(x_t^{1:2},u_t^{1:2})|\pi_t,\gamma_t^{1:2}) \quad (33)\\ &= \hat{c}(\pi_t,\gamma_t^{1:2})\end{aligned}$$

Proof: See Appendix D ■

Since $\{\Pi_t, t = 1, 2, ...T\}$ is a controlled Markov Process the optimum output functions can be given by $(\gamma_t^1, \gamma_t^2) = \psi_t(\pi_t)$. And thus optimum action by node k can be written as

$$u_t^k = g_t^k(x_t^k, \pi_t) \quad (34)$$

The dynamic program for the coordinator is

$$V(\pi) = \inf_{\gamma^{1:2}}[\hat{c}(\pi,\gamma^{1:2}) + \mathbb{E}\{\lambda V(\pi')|\pi,\gamma^{1:2}\}] \quad (35)$$

where the expectation is with respect to the conditional probability induced by the update function F and $u_t^{1:2}$ as random variable (noise). This result is in accordance with [4].

Furthermore, due to the specific nature of our problem, we show that instead of joint probability on the queue length of two nodes, individual marginals form a sufficient state. To that effect, we define random variable $\Xi_t^k \in \mathcal{P}(\mathbb{N})$ as the posterior pmf of X_t^k conditioned on $U_{1:t-1}^{1:2}, \Gamma_{1:t-1}^{1:2}$ i.e. $\Xi_t^k(x_t^k) = \mathbb{P}(X_t^k = x_t^k|U_{1:t-1}^{1:2}, \Gamma_{1:t-1}^{1:2})$ and show that (ξ_t^1, ξ_t^2) is controlled markov process. This gives a significant reduction in size of state over which optimum policies have to searched as π is defined over a space of $\mathcal{P}(\mathbb{N}^2)$ while (ξ^1, ξ^2) is defined over $\mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})$.

Lemma 5.

$$\xi_{t+1}^k = G^k(\xi_t^k, \gamma_t^k, u_t^{1:2}) \qquad k \in \{1,2\} \quad (36)$$

*where G^k is a deterministic update function that does not depend upon the policy **g***

Proof: For any fixed coordinator strategy ψ,

$$\xi^1_{t+1}(x^1_{t+1}) = \mathbb{P}^{\psi}(x^1_{t+1}|u^{1:2}_{1:t},\gamma^{1:2}_{1:t}) \tag{37a}$$

$$= \sum_{x^{1:2}_t} \mathbb{P}^{\psi}(x^1_{t+1}, x^{1:2}_t|u^{1:2}_{1:t},\gamma^{1:2}_{1:t}) \tag{37b}$$

$$= \sum_{x^{1:2}_t} \mathbb{P}^{\psi}(x^{1:2}_t|u^{1:2}_{1:t},\gamma^{1:2}_{1:t}).\mathbb{P}(x^1_{t+1}|x^1_t,u^{1:2}_t))\} \tag{37c}$$

Now,

$$\mathbb{P}^{\psi}(x^{1:2}_t|u^{1:2}_{1:t},\gamma^{1:2}_{1:t}) = \frac{\mathbb{P}^{\psi}(x^{1:2}_t,u^{1:2}_t|u^{1:2}_{1:t-1},\gamma^{1:2}_{1:t})}{\sum_{x^{1:2}_t}\mathbb{P}^{\psi}(x^{1:2}_t,u^{1:2}_t|u^{1:2}_{1:t-1},\gamma^{1:2}_{1:t})} \tag{37d}$$

$$= \frac{\mathbb{P}^{\psi}(x^{1:2}_t|u^{1:2}_{1:t-1},\gamma^{1:2}_{1:t})\mathbb{P}^{\psi}(u^{1:2}_t|u^{1:2}_{1:t-1},\gamma^{1:2}_{1:t},x^{1:2}_t)}{\sum_{x^{1:2}_t}\mathbb{P}^{\psi}(x^{1:2}_t,u^{1:2}_t|u^{1:2}_{1:t-1},\gamma^{1:2}_{1:t})} \tag{37e}$$

$$= \frac{\mathbb{P}^{\psi}(x^{1:2}_t|u^{1:2}_{1:t-1},\gamma^{1:2}_{1:t-1})\mathbf{1}_{u^{1:2}_t}\{\gamma^{1:2}_t(x^{1:2}_t)\}}{\sum_{x^{1:2}_t}\mathbb{P}^{\psi}(x^{1:2}_t|u^{1:2}_{1:t-1},\gamma^{1:2}_{1:t-1})\mathbf{1}_{u^{1:2}_t}\{\gamma^{1:2}_t(x^{1:2}_t)\}} \tag{37f}$$

$$\mathbb{P}^{\psi}(x^{1:2}_t|u^{1:2}_{1:t},\gamma^{1:2}_{1:t}) = \frac{\xi^1_t(x^1_t)\xi^2_t(x^2_t)\mathbf{1}_{u^{1:2}_t}\{\gamma^{1:2}_t(x^{1:2}_t)\}}{\sum_{x^{1:2}_t}\xi^1_t(x^1_t)\xi^2_t(x^2_t)\mathbf{1}_{u^{1:2}_t}\{\gamma^{1:2}_t(x^{1:2}_t)\}} \tag{37g}$$

Thus,

$$\xi^1_{t+1}(x^1_{t+1}) = \sum_{x^{1:2}_t} \frac{\xi^1_t(x^1_t)\xi^2_t(x^2_t)\mathbf{1}_{u^{1:2}_t}\{\gamma^{1:2}_t(x^{1:2}_t)\}}{\sum_{x^{1:2}_t}\xi^1_t(x^1_t)\xi^2_t(x^2_t)\mathbf{1}_{u^{1:2}_t}\{\gamma^{1:2}_t(x^{1:2}_t)\}}\mathbb{P}(x^1_{t+1}|x^1_t,u^{1:2}_t))\} \tag{37h}$$

$$= \sum_{x^1_t}\mathbb{P}(x^1_{t+1}|x^1_t,u^{1:2}_t)\frac{\xi^1_t(x^1_t)\mathbf{1}_{u^1_t}\{\gamma^1_t(x^1_t)\}\sum_{x^2_t}\mathbf{1}_{u^2_t}\{\gamma^2_t(x^2_t)\}\xi^2_t(x^2_t)}{\sum_{x^1_t}\xi^1_t(x^1_t)\mathbf{1}_{u^1_t}\{\gamma^1_t(x^1_t)\}\sum_{x^2_t}\mathbf{1}_{u^2_t}\{\gamma^2_t(x^2_t)\}\xi^2_t(x^2_t)} \tag{37i}$$

$$= \sum_{x^1_t}\mathbb{P}(x^1_{t+1}|x^1_t,u^{1:2}_t)\frac{\xi^1_t(x^1_t)\mathbf{1}_{u^1_t}\{\gamma^1_t(x^1_t)\}}{\sum_{x^1_t}\xi^1_t(x^1_t)\mathbf{1}_{u^1_t}\{\gamma^1_t(x^1_t)\}} \tag{37j}$$

$$= G^1(\xi^1_t,\gamma^1_t,u^{1:2}_t)(x^1_{t+1}) \tag{37k}$$

where (37g) is true since x^1_t and x^2_t are conditionally independent given $u^{1:2}_{t-1}$ (Lemma 2).

Similarly $\xi^2_{t+1} = G^2(\xi^2_t,\gamma^2_t,u^{1:2}_t)$ where G^1 and G^2 are deterministic functions.

■

Lemma 6. *The process* $\{(\Xi^1_t,\Xi^2_t); t = 1,2,...\}$ *is a controlled Markov Process with controls*

$\Gamma^{1:2}_t$. *i.e.*

$$\mathbb{P}(\xi^1_{t+1},\xi^2_{t+1}|\xi^1_{1:t},\xi^2_{1:t},\gamma^{1:2}_{1:t}) = \mathbb{P}(\xi^1_{t+1},\xi^2_{t+1}|\xi^1_t,\xi^2_t,\gamma^{1:2}_t) \tag{38}$$

$$\mathbb{E}(c(x^{1:2}_t,u^{1:2}_t)|\xi^1_{1:t},\gamma^{1:2}_{1:t}) = \hat{c}(\xi^1_t,\gamma^1_t)+\hat{c}(\xi^2_t,\gamma^2_t) \tag{39}$$

Proof: In the following we use the notation $G := (G^1, G^2)$

$$\mathbb{P}(\xi^1_{t+1},\xi^2_{t+1}|\xi^1_{1:t},\xi^2_{1:t},\gamma^{1:2}_{1:t}) = \sum_{u^{1:2}_t}\mathbb{P}(\xi^1_{t+1},\xi^2_{t+1},u^{1:2}_t|\xi^1_{1:t},\xi^2_{1:t},\gamma^{1:2}_{1:t}) \tag{40a}$$

$$= \sum_{u^{1:2}_t}\mathbf{1}_{\xi^1_{t+1},\xi^2_{t+1}}\{G(\xi^1_t,\xi^2_t,\gamma^{1:2}_t,u^{1:2}_t)\}\mathbb{P}(u^{1:2}_t|\xi^1_{1:t},\xi^2_{1:t},\gamma^{1:2}_{1:t}) \tag{40b}$$

$$= \sum_{u^{1:2}_t,x^{1:2}_t}\mathbf{1}_{\xi^1_{t+1},\xi^2_{t+1}}\{G(\xi_t,\xi^2_t,\gamma^{1:2}_t,u^{1:2}_t)\}\mathbf{1}_{u^{1:2}_t}\{\gamma^{1:2}_t(x^{1:2}_t)\}\mathbb{P}(x^{1:2}_t|\xi^1_{1:t},\xi^2_{1:t},\gamma^{1:2}_{1:t}) \tag{40c}$$

$$= \sum_{u^{1:2}_t,x^{1:2}_t}\xi^1_t(x^1_t)\xi^2_t(x^2_t)\mathbf{1}_{\xi^1_{t+1},\xi^2_{t+1}}\{G(\xi^1_t,\xi^2_t,\gamma^{1:2}_t,u^{1:2}_t)\}\mathbf{1}_{u^{1:2}_t}\{\gamma^{1:2}_t(x^{1:2}_t)\} \tag{40d}$$

$$= \sum_{x^{1:2}_t}\xi^1_t(x^1_t)\xi^2_t(x^2_t)\mathbf{1}_{\xi^1_{t+1},\xi^2_{t+1}}\{G(\xi^1_t,\xi^2_t,\gamma^{1:2}_t,\gamma^{1:2}_t(x^{1:2}_t))\} \tag{40e}$$

$$= \mathbb{P}(\xi^1_{t+1},\xi^2_{t+1}|\xi^1_t,\xi^2_t,\gamma^{1:2}_t) \tag{40f}$$

$$\mathbb{E}(c(x^{1:2}_t,u^{1:2}_t)|\xi^{1:2}_{1:t},\gamma^{1:2}_{1:t}) = \sum_{x^{1:2}_t,u^{1:2}_t}(x^1_t+x^2_t+u^1_t+u^2_t)\mathbb{P}(x^{1:2}_t,u^{1:2}_t|\xi^{1:2}_{1:t},\gamma^{1:2}_{1:t}) \tag{40g}$$

$$= \sum_{x^{1:2}_t,u^{1:2}_t}(x^1_t+x^2_t+u^1_t+u^2_t)\mathbb{P}(x^{1:2}_t|\xi^{1:2}_{1:t},\gamma^{1:2}_{1:t})\mathbf{1}_{u^{1:2}_t}\{\gamma^{1:2}_t(x^{1:2}_t)\} \tag{40h}$$

$$= \sum_{x^{1:2}_t,u^{1:2}_t}(x^1_t+x^2_t+u^1_t+u^2_t)\xi^1_t(x^1_t)\xi^2_t(x^2_t)\mathbf{1}_{u^{1:2}_t}\{\gamma^{1:2}_t(x^{1:2}_t)\} \tag{40i}$$

$$= \sum_{x^1_t}(x^1_t+\gamma^1_t(x^1_t))\xi^1_t(x^1_t)+\sum_{x^2_t}(x^2_t+\gamma^2_t(x^2_t))\xi^2_t(x^2_t) \tag{40j}$$

$$= \hat{c}(\xi^{1:2}_t,\gamma^{1:2}_t) \tag{40k}$$

■

where (40d) and (40i) are true because X^1_t, X^2_t are conditionally independent given $\Gamma^{1:2}_{1:t}, \Xi^{1:2}_{1:t}$ which are functions of $U^{1:2}_{1:t-1}$ (Lemma 2). Since $\{(\xi^1_t,\xi^2_t); t = 1, 2, ...\}$ is a controlled Markov Process, the optimum output functions can be given by $(\gamma^1_t,\gamma^2_t) = \psi_t(\xi^1_t,\xi^2_t)$. The dynamic programming equation can be given as

$$V(\xi^1,\xi^2) = \min_{\gamma^1,\gamma^2}[\hat{c}^1(\xi^1,\gamma^1) + \hat{c}^2(\xi^2,\gamma^2) + \lambda\mathbb{E}\{V(\underline{\xi^1},\underline{\xi^2})|\xi^{1:2},\gamma^{1:2}\}] \tag{41}$$

where the expectation is with respect to the conditional probability induced by the update functions (G^1, G^2) and $u_t^{1:2}$ as random variable (noise).

Thus finally, the optimum control actions are

$$u_t^k = g_t^k(x_t^k, \xi_t^1, \xi_t^2) \tag{42}$$

and action of node 1 is a function of its current queue length, its estimate of node 2's queue length (ξ_t^2) and also node 2's estimate of node 1's queue length (ξ_t^1).

This model cannot be extended to the case where there are more than two transmitter nodes (where the relay node in our model is also considered a transmitter node). This is because when there are only two transmitter nodes, in case of a collision, each node can determine that collision occurred due to simultaneous transmission of the other node. But if there are 3 or more nodes, in case of collision, a node cannot determine which other node(s) transmitted simultaneously. More precisely, in former case, feedback w_t combined with u_t^k gives transmission profile of each user i.e. (u_t^1, u_t^2) whereas in latter case, it is no longer true. Thus our model needs to be enriched so that each collision also contains the information regarding which nodes transmitted. This could be achieved if each node transmits a 'signature' waveform along with the data waveform such that signature waveform of all users are mutually orthogonal and orthogonal to data (for e.g. in frequency).

V. Conclusion and Future work

We analyzed energy delay tradeoff in a simple relay channel. In section III we found the optimum centralized if relay does not have any traffic and showed that it can be implemented in a decentralized way also. In section IV we proved the structural result that the optimum policy can be found using solving a dynamic programming equation. The domain of optimization is space of probability mass functions on state space $\mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})$ and is still intractable. Future work can analyze more structural properties of the optimum strategy to design optimum or suboptimum strategies and analyze its performance.

APPENDIX

A. List of symbols

$$K = \frac{\lambda p}{1 - \lambda + \lambda p}$$

$$c = \frac{E_{13}(1-\lambda) - \lambda(1-p)}{(1-\lambda)^2}$$

$$s = \frac{1}{1-\lambda}$$

$$d = \frac{\lambda(1-\lambda p) - E_{13}(1-\lambda)(1-\lambda p)}{(1-\lambda)^2}$$

$$\rho = \frac{\lambda - \lambda p}{1 - \lambda p}$$

$$\alpha = \frac{(\lambda p)^2}{1 - 2\lambda^2 p(1-p)}$$

$$\beta = \frac{(\lambda(1-p))^2}{1 - 2\lambda^2 p(1-p)}$$

$$\theta = \frac{1+\lambda}{1 - 2\lambda^2 p(1-p)}$$

$$\gamma = \frac{E_{12} + \lambda E_{23} + \lambda p}{1 - 2\lambda^2 p(1-p)}$$

$$\xi = a\beta$$

$$\eta = \frac{a\gamma}{(1-a\alpha)} + \frac{a^2\alpha\theta}{(1-a\alpha)^2}$$

$$\omega = \frac{a\theta}{(1-a\alpha)}$$

$$u = \frac{\eta+\omega}{1-\xi} - \frac{\omega}{(1-\xi)^2}$$

$$= \frac{-\lambda^2(1 + E23 - p) + \lambda(E_{23} + p - E_{12}) + E_{12}}{(1-\lambda)^2(1+\lambda)}$$

$$r = V(0) - u = \frac{Ks + Ku - u}{1 - K\xi}$$

$$l = \left[(u-c)(1-\rho) + (\rho - a\beta)(u + s(x_{th} - 1) - V(x_{th} - 1))\right]$$

$$w_1(x) = \left[(a-1)^{x-1}\left(\frac{1}{(1-K\xi)} - \frac{1}{2-a}\right) + \frac{1}{2-a}\right]$$

$$a_{x_{th}} = a\alpha(a\alpha - a + 1)\left(\frac{1}{1-Ka\beta} - \frac{1}{2-a}\right)(a-1)^{xth-2} + \frac{(1-a\alpha\rho)(1-a\alpha)}{(2-a)(\rho-\xi)}$$

$$b_{x_{th}} = \frac{(1-\lambda p)(a\beta)^{x_{th}-1}(1-K)(1-a\alpha\rho)}{(1-\lambda^2)(1-Ka\beta)} + \frac{a_{x_{th}}}{1+\lambda}$$

B. Conditions sufficient for uniqueness of the optimum solution

Let $S = \{0, 1, 2, \cdots\}$ and $w : S \to \mathbb{R}$ be a real valued function such that $\inf_{s \in S} w(s) > 0$ and induce weighted supremum norm for real valued functions v on S as $\|v\|_w = \sup_{s \in S} w(s)^{-1}|v(s)|$. Let V_w be the space of real valued functions v on S satisfying $\|v\|_w < \infty$. Then V_w is a Banach space (complete normed linear space) and convergence in V_w with respect to weighted sup norm implies point wise convergence, since $\|v^n - v\|_w < \epsilon$ implies $|v^n(s) - v(s)| < \epsilon w(s)$

Let $c(x, a)$ be the instantaneous cost where x is the state and a is the action , $a \in A_x$, A_x is the set os possible actions with state x, $P(j|x, a)$ be the transition probability from state x to state j under action a and $P_\pi^J(j|x)$ be probability of reaching state j from state x in J steps with policy $\pi = (g_1, \cdots g_J)$. [3, sec. 6.10] shows that if following conditions are satisfied, then the optimal dynamic programming equation has a unique solution in V_w which is equal to the optimal cost.

1) There exists a constant $\mu < \infty$ such that

$$\sup_{a \in A_x} |r(x, a)| \leq \mu w(x)$$

2) (a) There exists constant k, $0 \leq k < \infty$ such that for all $a \in A_x$, $x \in S$

$$\sum_{j \in S} P(j|x, a) w(j) \leq k w(x)$$

(b) For each λ, $0 \leq \lambda < 1$, there exists α , $0 \leq \alpha < 1$ and an integer J such that for Markov policies $\pi = (g_1, \cdots g_J)$

$$\lambda^J \sum_{j \in S} P_\pi^J(j|x) w(j) \leq \alpha w(x)$$

C. Finding Cost-to-go function

In general, V(x) of the form $V(x+1) = AV(x) + Bx + C$ can be solved as follows

$$\begin{aligned}
V(x+1) &= AV(x) + Bx + C \\
V(x+1) &= A(AV(x-1) + B(x-1) + C) + Bx + C \\
&= A^2V(x-1) + AB(x-1) + AC + Bx + C \\
&= A^2V(x-1) + B(x + A(x-1)) + C(1+A) \\
&= A^{x+1}V(0) + B(x + A(x-1) + ... + A^{x-1}) + C(1 + A + ... + A^x) \\
&= A^{x+1}V(0) + B\left(\frac{x}{1-A} - \frac{A(1-A^x)}{(1-A)^2}\right) + C\frac{(1-A^{x+1})}{(1-A)} \\
V(x+1) &= (x+1)\frac{B}{1-A} + \left(\frac{C}{1-A} - \frac{B}{(1-A)^2}\right) + A^{x+1}\left(V(0) + \frac{B}{(1-A)^2} - \frac{C}{(1-A)}\right)
\end{aligned}$$

1) Cost-to-go function for waiting: We find the cost-to-go function when the policy is to wait for all x.

$$\begin{aligned}
V(x) &= x + \lambda p V(x+1) + \lambda(1-p)V(x) \\
V(x+1) &= \frac{1-\lambda(1-p)}{\lambda p}V(x) - \frac{1}{\lambda p}x
\end{aligned}$$

Here $A = \frac{1-\lambda(1-p)}{\lambda p}, B = -\frac{1}{\lambda p}, C = 0$ which gives,

$$V(x) = \left(V(0) - \frac{\lambda p}{(1-\lambda)^2}\right)\left[\frac{1-\lambda(1-p)}{\lambda p}\right]^x + \frac{\lambda p}{(1-\lambda)^2} + \frac{1}{1-\lambda}x \tag{43}$$

The above $V(x)$ satisfies the recursive equation for all $V(0)$.

For uniqueness (appendix B), let $w(x) = x + 1 + E_{12} + E_{13} + E_{23}$. Then $w(x) \geq 1$ and

1)

$$\sup_{a \in A_x} |c(x,a)| \leq w(x)$$

2) for all $a \in A_x$, $x \in S$

$$\sum_{j \in S} P(j|x,a)w(j) \leq w(x+1) \leq 2w(x)$$

3) Since in each state, there are only at most one step transitions possible, thus

$$\sum_{j\in S} P_\pi^J(j|x)w(j) \leq w(J+x)$$

and for $\lambda \in [0,1)$, since $w(x)$ is linear (affine) in x, there exists a J integer and there exists α , $0 \leq \alpha < 1$ such that

$$\lambda^J \sum_{j\in S} P_\pi^J(j|x)w(j) \leq \lambda^J w(J+x) \leq \alpha w(x)$$

Thus there exists a unique $V \in V_w$ such that V is the unique solution of the dynamic programming equation and is equal to the optimum cost.

For $V \in V_w, \sup_x |V(x)w^{-1}(x)| < \infty$, which implies $V(0) = \frac{\lambda p}{(1-\lambda)^2}$ since $\frac{1-\lambda(1-p)}{\lambda p} > 1$
Thus

$$V(x) = \frac{\lambda p}{(1-\lambda)^2} + \frac{1}{1-\lambda}x \tag{44}$$

2) Cost-to-go when transmitting directly : The cost-to-go function in the recursive form is given as

$$\begin{aligned} V(0) &= \lambda p V(1) + \lambda(1-p)V(0) \\ V(x) &= x + E_{13} + \lambda p V(x) + \lambda(1-p)V(x-1) \qquad (x > 0) \end{aligned}$$

Here $A = \frac{\lambda(1-p)}{1-\lambda p}, B = \frac{1}{1-\lambda p}, C = \frac{E_{13}}{1-\lambda p}$ which gives

$$V(x+1) = \frac{(x+1)}{1-\lambda} + \frac{E_{13}(1-\lambda) - \lambda(1-p)}{(1-\lambda)^2} + [\frac{\lambda(1-p)}{1-\lambda p}]^{x+1}\left(V(0) - \frac{E_{13}(1-\lambda)-\lambda(1-p)}{(1-\lambda)^2}\right)$$

Calculating expression for V(1) from this,

$$\begin{aligned} V(1) &= \frac{1}{1-\lambda} + \frac{E_{13}(1-\lambda) - \lambda(1-p)}{(1-\lambda)^2} + [\frac{\lambda(1-p)}{1-\lambda p}]^{1}\left(V(0) - \frac{E_{13}(1-\lambda)-\lambda(1-p)}{(1-\lambda)^2}\right) \\ V(0) &= \lambda p V(1) + \lambda(1-p)V(0) \end{aligned}$$

This gives

$$V(x) = c + sx + d\rho^x \tag{45}$$

where

$$\begin{array}{rlrl} c = & \frac{E_{13}(1-\lambda)-\lambda(1-p)}{(1-\lambda)^2} & s = & \dfrac{1}{1-\lambda} \\ d = & \frac{\lambda(1-\lambda p)-E_{13}(1-\lambda)(1-\lambda p)}{(1-\lambda)^2} & \rho = & \dfrac{\lambda-\lambda p}{1-\lambda p} \end{array}$$

3) Cost-to-go Function when transmitting through relay : We find the cost-to-go function of the policy where successful transmission of a packet from node 1 to node 3 is achieved through node 2. If there is a packet waiting in the queue of node 1 and there is no packet in the queue of node 2, then node 1 transmits the packet to node 2 in next slot and then the node 2 transmits the packet to node 3 in the subsequent slot while node 1 waits in this slot.

$$\begin{aligned} V(0) = &\quad \lambda p V(1) + \lambda(1-p)V(0) \\ V(0) = &\quad \frac{\lambda p}{1-\lambda+\lambda p} V(1) \\ \text{Let } K = &\quad \frac{\lambda p}{1-\lambda+\lambda p} \\ V(0) = &\quad K V(1) \end{aligned}$$

$\forall x \geq 1$, the general equation is given by

$$\begin{aligned} V(x) &= x(1+\lambda) + (E_{12} + \lambda E_{23} + \lambda p) + \lambda^2 p^2 V(x+1) + \lambda^2(1-p)^2 V(x-1) + 2\lambda^2 p(1-p)V(x) \\ V(x) &= \gamma + \theta x + \alpha V(x+1) + \beta V(x-1) \end{aligned}$$

where,

$$\begin{array}{rlrl} \alpha = & \frac{(\lambda p)^2}{1-2\lambda^2 p(1-p)} & \beta = & \dfrac{(\lambda(1-p))^2}{1-2\lambda^2 p(1-p)} \\ \theta = & \frac{1+\lambda}{1-2\lambda^2 p(1-p)} & \gamma = & \dfrac{E_{12}+\lambda E_{23}+\lambda p}{1-2\lambda^2 p(1-p)} \end{array}$$

$$\begin{aligned} V(x) &= \gamma + \theta x + \alpha V(x+1) + \beta V(x-1) \\ \begin{bmatrix} V(x) \\ V(x-1) \end{bmatrix} &= \begin{bmatrix} 0 & 1 \\ -\frac{\alpha}{\beta} & \frac{1}{\beta} \end{bmatrix} \begin{bmatrix} V(x+1) \\ V(x) \end{bmatrix} + \begin{bmatrix} 0 \\ -\frac{\theta}{\beta} \end{bmatrix} x + \gamma \begin{bmatrix} 0 \\ -\frac{1}{\beta} \end{bmatrix} \end{aligned}$$

$$\mathbf{W}(x) = \begin{bmatrix} V(x) \\ V(x-1) \end{bmatrix} \ \mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\frac{\alpha}{\beta} & \frac{1}{\beta} \end{bmatrix} \ \mathbf{B} = \begin{bmatrix} 0 \\ -\frac{\theta}{\beta} \end{bmatrix} \ \mathbf{C} = \begin{bmatrix} 0 \\ -\frac{1}{\beta} \end{bmatrix}$$

$$\begin{aligned} \mathbf{W}(x) &= \mathbf{A}\mathbf{W}(x+1) + \gamma\mathbf{C} + \mathbf{B}x \\ &= \mathbf{A}^n\mathbf{W}(x+n) + \gamma(1 + \mathbf{A} + ... + \mathbf{A}^{n-1})\mathbf{C} + (x + \mathbf{A}(x+1) + ... + \mathbf{A}^{n-1}(x+n-1))\mathbf{B} \end{aligned}$$

Eigenvalue decomposition of A gives,

$$\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^{-1} \quad \text{where} \quad \mathbf{U} = \begin{bmatrix} \frac{1}{\nu_1} & \frac{1}{\nu_2} \\ 1 & 1 \end{bmatrix}$$

$$\nu_1 = \tfrac{1-\sqrt{1-4\alpha\beta}}{2\beta}(\nu_1 < 1) \quad \nu_2 = \tfrac{1+\sqrt{1-4\alpha\beta}}{2\beta}(\nu_2 > 1)$$

$$\mathbf{A}^n = \mathbf{U}\Lambda^n\mathbf{U}^{-1}$$

This gives,

$$\frac{\nu_1\nu_2}{\nu_2 - \nu_1} \begin{bmatrix} V(x) - \frac{V(x-1)}{\nu_2} \\ -V(x) + \frac{V(x-1)}{\nu_1} \end{bmatrix} = \begin{bmatrix} \nu_1^n & 0 \\ 0 & \nu_2^n \end{bmatrix} \mathbf{U}^{-1}\mathbf{W}(x+n) + \gamma \begin{bmatrix} \frac{1-\nu_1^n}{1-\nu_1} & 0 \\ 0 & \frac{1-\nu_2^n}{1-\nu_2} \end{bmatrix} \mathbf{U}^{-1}\mathbf{C} +$$
$$\begin{bmatrix} \frac{x}{1-\nu_1} + \frac{\nu_1}{(1-\nu_1)^2} - \nu_1^n(\frac{(x+n-1)}{1-\nu_1} + \frac{1}{(1-\nu_1)^2}) & 0 \\ 0 & \frac{x}{1-\nu_2} + \frac{\nu_2}{(1-\nu_2)^2} - \nu_2^n(\frac{(x+n-1)}{1-\nu_2} + \frac{1}{(1-\nu_2)^2}) \end{bmatrix} \mathbf{U}^{-1}\mathbf{B}$$

Now we are interested in $\nu_1 < 1$, since that results in $V \in V_w$

$$\begin{aligned} \frac{\nu_1\nu_2}{\nu_2 - \nu_1} \begin{bmatrix} V(x) - \frac{V(x-1)}{\nu_2} \end{bmatrix} &= \begin{bmatrix} \nu_1^n & 0 \end{bmatrix} \mathbf{U}^{-1}\mathbf{W}(x+n) + \gamma \begin{bmatrix} \frac{1-\nu_1^n}{1-\nu_1} & 0 \end{bmatrix} \mathbf{U}^{-1}\mathbf{C} \\ &+ \begin{bmatrix} \frac{x}{1-\nu_1} + \frac{\nu_1}{(1-\nu_1)^2} - \nu_1^n(\frac{(x+n-1)}{1-\nu_1} + \frac{1}{(1-\nu_1)^2}) & 0 \end{bmatrix} \mathbf{U}^{-1}\mathbf{B} \end{aligned}$$

Taking limit n to ∞ given that $\nu_1 < 1$

$$V(x) \;=\; \frac{1}{\nu_2}V(x-1) + x\frac{\theta}{(1-\nu_1)\beta\nu_2} + \frac{1}{\beta\nu_2(1-\nu_1)}(\gamma + \frac{\nu_1\theta}{(1-\nu_1)})$$

Let $\nu_1 = a\alpha, \nu_2 = \frac{1}{a\beta}$ so that $a = \frac{1-\sqrt{1-4\alpha\beta}}{2\alpha\beta}$ and a satisfies the equation $(\alpha\beta)a^2 - a + 1 = 0$

$$V(x) = \xi V(x-1) + \omega x + \eta$$

where

$$\xi = a\beta \quad \eta \;= \tfrac{a\gamma}{(1-a\alpha)} + \tfrac{a^2\alpha\theta}{(1-a\alpha)^2} \quad \omega = \tfrac{a\theta}{(1-a\alpha)}$$

$$\begin{aligned} V(1) &= \xi V(0) + \omega + \eta \\ V(0) &= KV(1) \\ V(0) &= \frac{K(\eta+\omega)}{1-K\xi} \end{aligned}$$

Let

$$u = \tfrac{\eta+\omega}{1-\xi} - \tfrac{\omega}{(1-\xi)^2} = \tfrac{-\lambda^2(1+E23-p)+\lambda(E_{23}+p-E_{12})+E_{12}}{(1-\lambda)^2(1+\lambda)} \quad r = V(0) - u = \tfrac{Ks+Ku-u}{1-K\xi} \quad s = \tfrac{\omega}{1-\xi} = \tfrac{1}{1-\lambda}$$

Thus

$$V(x) = r\xi^x + sx + u \tag{46}$$

D. Decentralized control

Proposition 2.

$$\pi_{t+1} = F(\pi_t, \gamma_t^{1:2}, u_t^{1:2})$$

where F is the update function that does not depend upon the policy ***g***

Proof: Fix ψ

$$\begin{aligned}
\pi_{t+1}(x_{t+1}^{1:2}) &= \mathbb{P}(X_{t+1} = x_{t+1}^{1:2}|u_{1:t}^{1:2}, \gamma_{1:t}^{1:2}) \\
&= \sum_{x_t^{1:2}} \mathbb{P}(x_{t+1}^{1:2}, x_t^{1:2}|u_{1:t}^{1:2}, \gamma_{1:t}^{1:2}) \\
&= \sum_{x_t^{1:2}} \mathbb{P}(x_t^{1:2}|u_{1:t}^{1:2}, \gamma_{1:t}^{1:2}).\mathbb{P}(x_{t+1}^{1:2}|x_t^{1:2}, u_t^{1:2})
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{P}(x_t^{1:2}|u_{1:t}^{1:2}, \gamma_{1:t}^{1:2}) &= \frac{\mathbb{P}(x_t^{1:2}, u_t^{1:2}|u_{1:t-1}^{1:2}, \gamma_{1:t}^{1:2})}{\sum_{x_t^{1:2}} \mathbb{P}(x_t^{1:2}, u_t^{1:2}|u_{1:t-1}^{1:2}, \gamma_{1:t}^{1:2})} \\
&= \frac{\mathbb{P}(x_t^{1:2}|u_{1:t-1}^{1:2}, \gamma_{1:t}^{1:2})\mathbb{P}(u_t^{1:2}|u_{1:t-1}^{1:2}, \gamma_{1:t}^{1:2}, x_t^{1:2})}{\sum_{x_t^{1:2}} \mathbb{P}(x_t^{1:2}, u_t^{1:2}|u_{1:t-1}^{1:2}, \gamma_{1:t}^{1:2})} \\
&= \frac{\mathbb{P}(x_t^{1:2}|u_{1:t-1}^{1:2}, \gamma_{1:t-1}^{1:2})\mathbf{1}_{u_t^{1:2}}\{\gamma_t^{1:2}(x_t^{1:2})\}}{\sum_{x_t^{1:2}} \mathbb{P}(x_t^{1:2}|u_{1:t-1}^{1:2}, \gamma_{1:t-1}^{1:2})\mathbf{1}_{u_t^{1:2}}\{\gamma_t^{1:2}(x_t^{1:2})\}}
\end{aligned}$$

Since $\gamma_t^{1:2} = \psi(u_{1:t-1}^{1:2})$

$$\mathbb{P}(x_t^{1:2}|u_{1:t}^{1:2}, \gamma_{1:t}^{1:2}) = \frac{\pi_t^{1:2}(x_t^{1:2})\mathbf{1}_{u_t^{1:2}}\{\gamma_t^{1:2}(x_t^{1:2})\}}{\sum_{x_t^{1:2}} \pi_t^{1:2}(x_t^{1:2})\mathbf{1}_{u_t^{1:2}}\{\gamma_t^{1:2}(x_t^{1:2})\}}$$

Thus,

$$\pi_{t+1} = F(\pi_t, \gamma_t^{1:2}, u_t^{1:2})$$

■

Lemma 7. *The process* $\{\Pi_t, t = 1, 2, ...T\}$ *is a controlled Markov Process with control* $\gamma_t^{1:2}$*. i.e.*

$$\begin{aligned}
\mathbb{P}(\pi_{t+1}|\pi_{1:t}, \gamma_{1:t}^{1:2}) &= \mathbb{P}(\pi_{t+1}|\pi_t, \gamma_t^{1:2}) \\
\mathbb{E}(c(x_t^{1:2}, u_t^{1:2})|\pi_{1:t}, \gamma_{1:t}^{1:2}) &= \hat{c}(\pi_t, \gamma_t^{1:2})
\end{aligned}$$

Proof:

$$\begin{aligned}
\mathbb{P}(\pi_{t+1}|\pi_{1:t},\gamma_{1:t}^{1:2}) &= \sum_{u_t^{1:2}} \mathbb{P}(\pi_{t+1},u_t^{1:2}|\pi_{1:t},\gamma_{1:t}^{1:2}) \\
&= \sum_{u_t^{1:2}} \mathbf{1}_{\pi_{t+1}}\{F(\pi_t,\gamma_t^{1:2},u_t^{1:2})\}\mathbb{P}(u_t^{1:2}|\pi_{1:t},\gamma_{1:t}^{1:2}) \\
&= \sum_{u_t^{1:2},x_t^{1:2}} \mathbf{1}_{\pi_{t+1}}\{F(\pi_t,\gamma_t^{1:2},u_t^{1:2})\}\mathbf{1}_{u_t^{1:2}}\{\gamma_t^{1:2}(x_t^{1:2})\}\mathbb{P}(x_t^{1:2}|\pi_{1:t},\gamma_{1:t}^{1:2}) \\
&= \sum_{u_t^{1:2},x_t^{1:2}} \pi_t(x_t^{1:2})\mathbf{1}_{\pi_{t+1}}\{F(\pi_t,\gamma_t^{1:2},u_t^{1:2})\}\mathbf{1}_{u_t^{1:2}}\{\gamma_t^{1:2}(x_t^{1:2})\} \\
&= \mathbb{P}(\pi_{t+1}|\pi_t,\gamma_t^{1:2})
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(c(x_t^{1:2},u_t^{1:2})|\pi_{1:t},\gamma_{1:t}^{1:2}) &= \sum_{x_t^{1:2},u_t^{1:2}} c(x_t^{1:2},u_t^{1:2})\mathbb{P}(x_t^{1:2},u_t^{1:2}|\pi_{1:t},\gamma_{1:t}^{1:2}) \\
&= \sum_{x_t^{1:2},u_t^{1:2}} c(x_t^{1:2},u_t^{1:2})\mathbb{P}(x_t^{1:2}|\pi_{1:t},\gamma_{1:t}^{1:2})\mathbf{1}_{u_t^{1:2}}\{\gamma_t^{1:2}(x_t^{1:2})\} \\
&= \sum_{x_t^{1:2},u_t^{1:2}} c(x_t^{1:2},u_t^{1:2})\pi_t(x_t^{1:2})\mathbf{1}_{u_t^{1:2}}\{\gamma_t^{1:2}(x_t^{1:2})\} \\
&= \hat{c}(\pi_t,\gamma_t^{1:2})
\end{aligned}$$

■

References

[1] P. R. Kumar and P. Varaiya , “Stochastic systems: estimation, identification, and adaptive control”, Englewood Cliffs, NJ: Prentice-Hall, 1986.

[2] D. Bertsekas , “Dynamic Programming and Stochastic Control”, Academic Press, 1976.

[3] Martin L. Puterman , “ Markov Decision Processes: Discrete Stochastic Dynamic Programming”, Wiley, 1994.

[4] A. Nayyar, A. Mahajan and D.Teneketzis, “Optimal Control Strategies in Delayed Sharing Information Structures”, *IEEE Transactions on Automatic Control*, ,Oct. 2010