# The Average Waiting Time for Both Classes in a Delayed Accumulating Priority Queue

Blair Bilodeau[1] and David Stanford[2]

[1]University of Toronto, Department of Statistical Sciences
[2]Western University, Department of Statistical and Actuarial Sciences

May 27, 2019

Presented to the Canadian Operational Research Society in Saskatoon, Canada

# Overview

# Accumulating Priority Queue

**Problem Formulation**

Class-1 and Class-2 customers arrive with zero priority.

- Arrival rates: $\lambda_1, \lambda_2 \in [0, \infty)$
- Priority accumulation rates: $b_1 > b_2 \in (0, \infty)$
- Service rate: $\mu \in (0, \infty)$
- Stability: $\rho := \frac{\lambda_1 + \lambda_2}{\mu} < 1$

**Accumulated Priority**

Consider the $n^{\text{th}}$ customer, of class $i(n)$, who arrived at $\tau_n$:

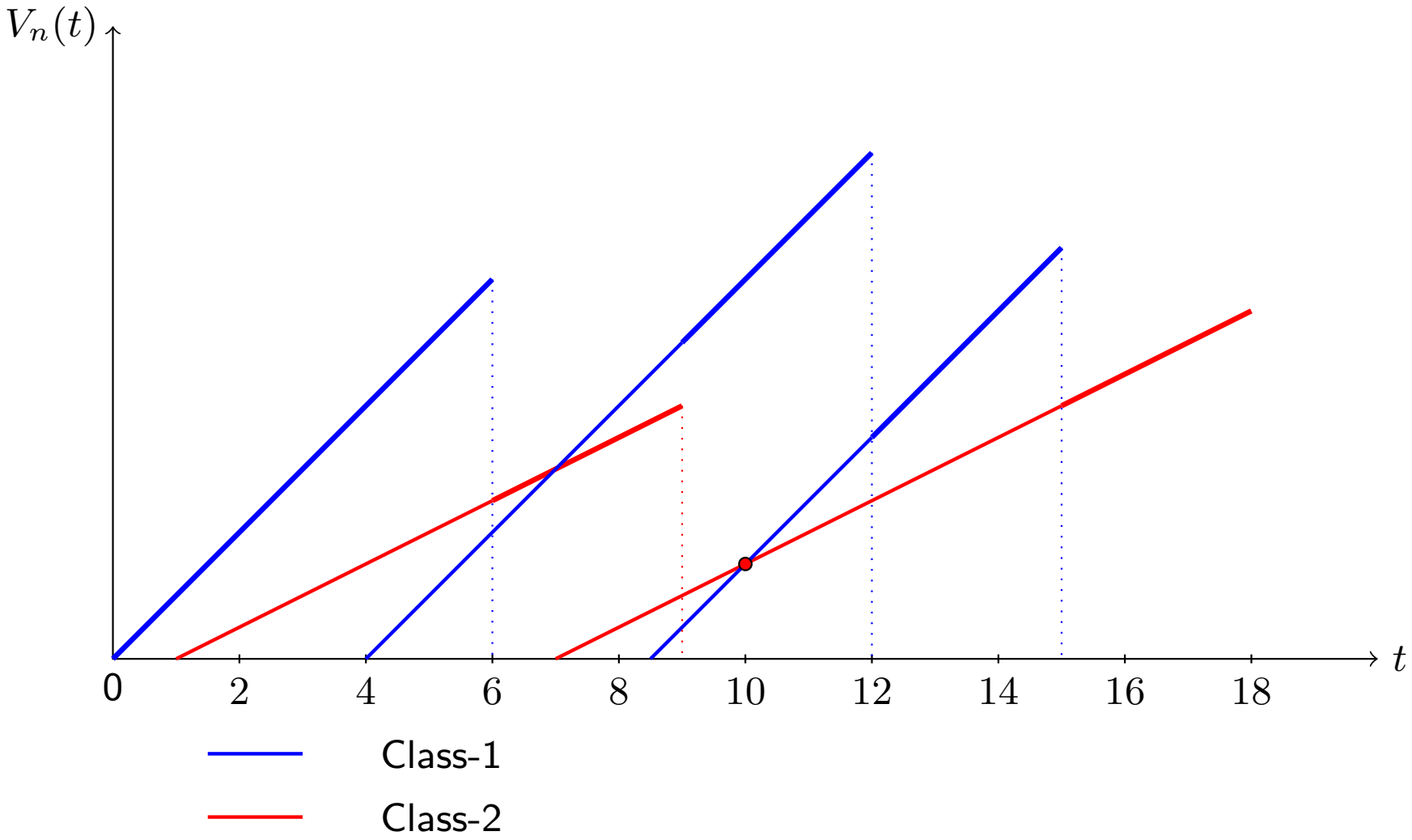
$$V_n(t) = b_{i(n)}(t - \tau_n)$$

**Limitation**

Heavily penalizes Class-1 compared to Non-Preemptive Priority Queue.

# Accumulating Priority Queue

A sample of accumulated priority in an Accumulating Priority Queue:



# Delayed Accumulating Priority Queue

**Motivation**

In hospital settings, some patients may not need to be seen urgently until after some time has passed. This allows more preference to be given to Class-1 customers, while not ignoring Class-2.

**Additional Structure**

- For simplicity, $b_1 = 1$ and $b_2 := b \in (0, 1)$
- Class-2 waits for $d \in (0, \infty)$ units of time before accumulating priority

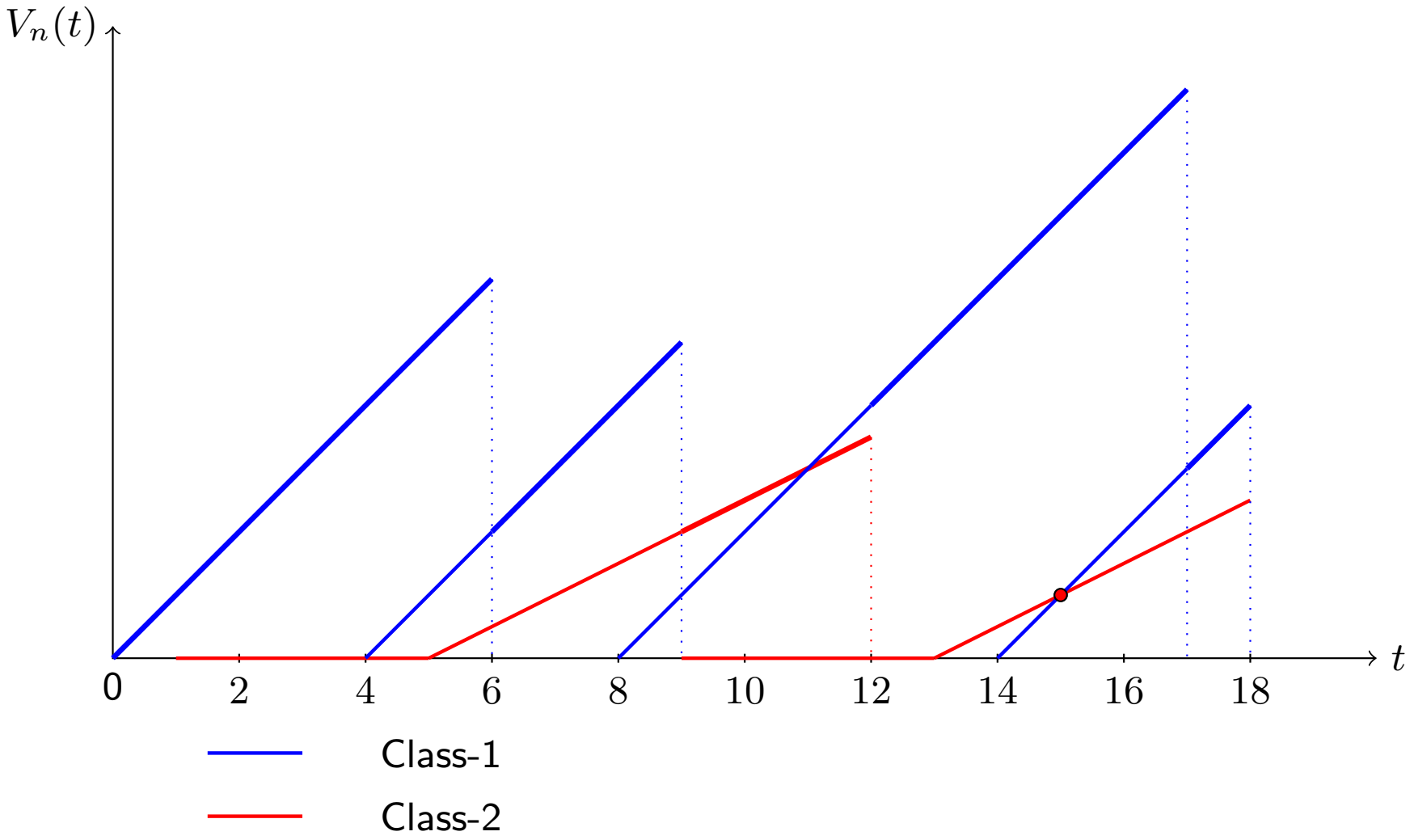
**Accumulated Priority**

Consider the $n^{\text{th}}$ customer, of class $i(n)$, who arrived at $\tau_n$:

$$V_n(t) = \begin{cases} t - \tau_n & \text{if} \quad i(n) = 1 \\ b(t - d - \tau_n) & \text{if} \quad i(n) = 2 \end{cases}$$

# Delayed Accumulating Priority Queue

A sample of accumulated priority in a Delayed APQ:



# Waiting Time Less Than $d$

Let $\mathcal{W}_{DAP,i}$ and $\mathcal{W}_{NP,i}$ denote the stationary waiting time for Class-$i$ patients from the Delayed APQ and Non-Preemptive Priority Queue respectively.

# Waiting Time Less Than $d$

Let $\mathcal{W}_{DAP,i}$ and $\mathcal{W}_{NP,i}$ denote the stationary waiting time for Class-$i$ patients from the Delayed APQ and Non-Preemptive Priority Queue respectively.

**Theorem 3.1. (Mojalal et al. 2019)**

*Up to time $d$, the waiting time for a Class-2 customer in the Delayed APQ is the same as the waiting time for a Class-2 customer in the Non-Preemptive priority queue. That is,*

$$P(\mathcal{W}_{DAP,2} \leq t) = P(\mathcal{W}_{NP,2} \leq t) \ \forall t \in [0, d].$$

**Implication**

Since the distribution of $\mathcal{W}_{NP,2}$ is known, we only have to consider the case when waiting is longer than $d$ units of time.

# Waiting Time Greater Than $d$

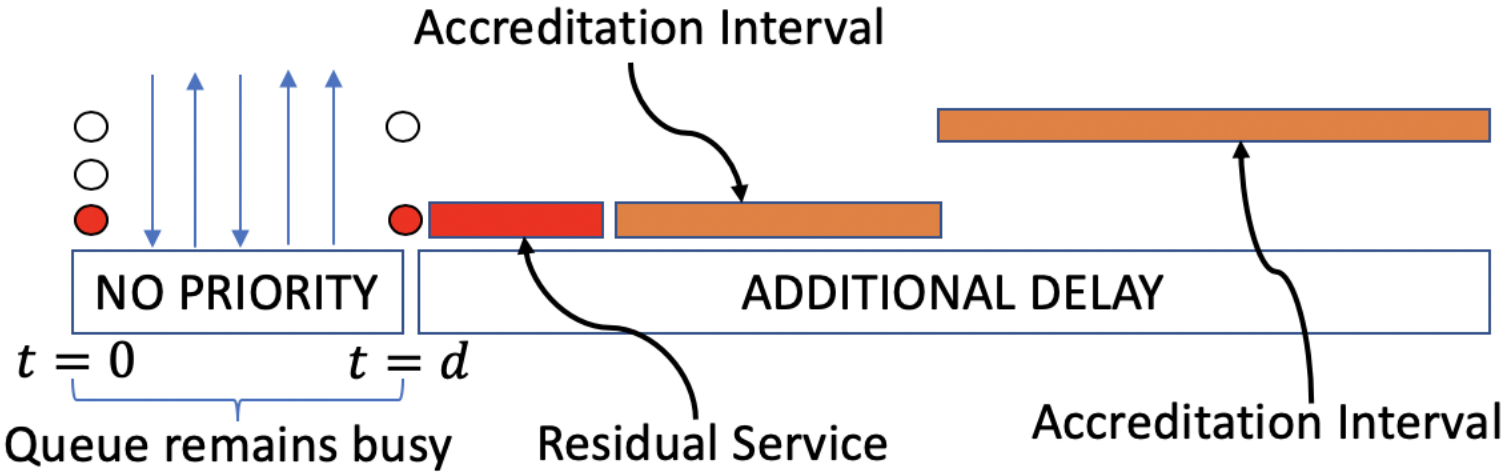
**Strategy**

Consider a Class-2 customer of interest, denoted by $X$, who has been waiting for $d$ units of time. The following determine the waiting time of $X$ beyond $d$:

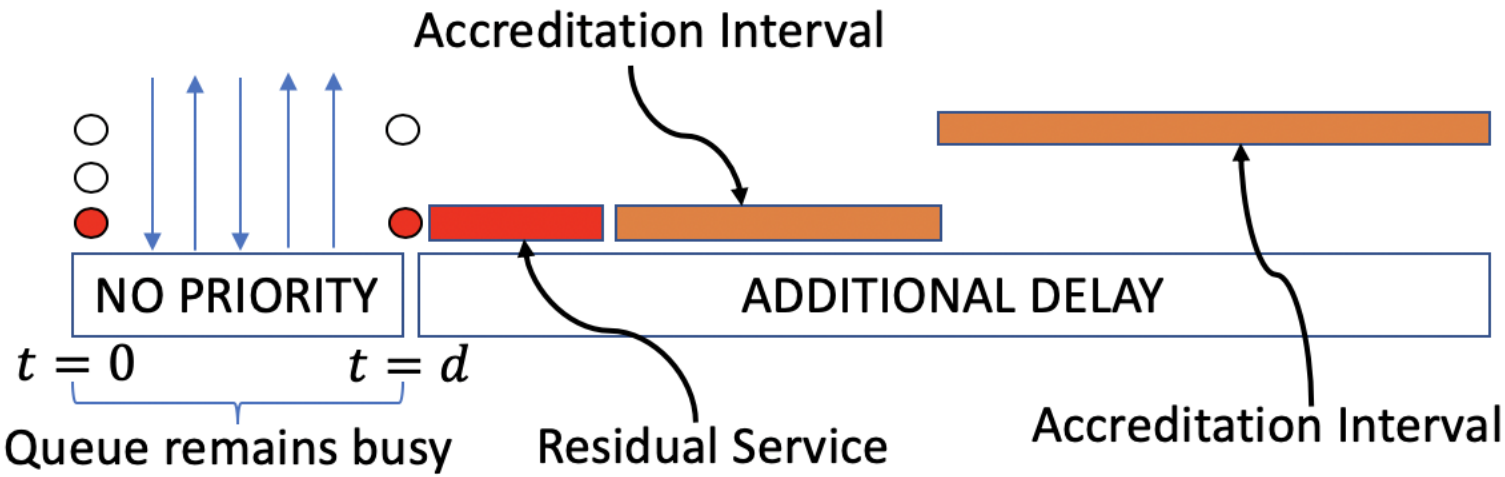
1) The customer currently in service must finish service.
2) All customers in the system at time $d$ with greater priority than $X$ must be served.
3) All customers who accumulate more priority than $X$ before $X$ enters service must be served. These are referred to as *accrediting* customers.

Each customer generates an *accreditation interval* consisting of their service time plus the service times of all those who accredit during their service.

# Accreditation Intervals



# Accreditation Intervals




**Length of an Accreditation Interval**

The distribution of the length of an accreditation interval is completely determined by the rate at which customers accredit.

- Delayed APQ: $\lambda_1(1-b)$
- Non-Preemptive Priority Queue: $\lambda_1$

**Intuition**

The reduced waiting time experienced by Class-2 customers in the Delayed APQ can be completely explained by the lower accreditation rate, and consequently shorter accreditation intervals.

# Waiting Time Greater Than $d$

Let $N_t$ be the number of customers in system $t$ time units after arrival.

- $\pi_i := P(N_0 = i)$
- $P_{ij}(d) := P\left(N_d = j, N_t > 0; t \in [0, d] \mid N_0 = i\right)$

Denote the Laplace-Stieltjes transform of an accreditation interval for queue type $Q$ by $\eta_Q(s)$.

# Waiting Time Greater Than $d$

Let $N_t$ be the number of customers in system $t$ time units after arrival.

- $\pi_i := P(N_0 = i)$
- $P_{ij}(d) := P\left(N_d = j, N_t > 0; t \in [0, d] \mid N_0 = i\right)$

Denote the Laplace-Stieltjes transform of an accreditation interval for queue type $Q$ by $\eta_Q(s)$.

## Theorem 3.2. (Mojalal et al. 2019)

*In the M/M/1 case, The LST of the waiting time greater than $d$ is*

$$E\left[e^{-s\mathcal{W}_{Q,2}}\mathbb{1}\{\mathcal{W}_{Q,2} > d\}\right] = \sum_{i=1}^{\infty}\pi_i \sum_{j=1}^{\infty} P_{ij}(d)e^{-sd}\left(\eta_Q(s)\right)^j,$$

$$Q \in \{DAP, NP\}.$$

# Class-2 M/M/1 Average Waiting Time

**Removing the Laplace-Stieltjes Transform**

$$
\begin{aligned}
& E\left[\mathcal{W}_{Q,2}\mathbb{1}\{\mathcal{W}_{Q,2} > d\}\right] \\
= & -\frac{d}{ds}E\left[e^{-s\mathcal{W}_{Q,2}}\mathbb{1}\{\mathcal{W}_{Q,2} > d\}\right]\Big|_{s=0} \\
= & \sum_{i=1}^{\infty}\pi_i\sum_{j=1}^{\infty}P_{ij}(d)\left[d + j\eta_Q'(0)\right].
\end{aligned}
$$

# Class-2 M/M/1 Average Waiting Time

## Removing the Laplace-Stieltjes Transform

$$\begin{aligned} & E\left[\mathcal{W}_{Q,2}\mathbb{1}\{\mathcal{W}_{Q,2} > d\}\right] \\ = & -\frac{d}{ds}E\left[e^{-s\mathcal{W}_{Q,2}}\mathbb{1}\{\mathcal{W}_{Q,2} > d\}\right]\Big|_{s=0} \\ = & \sum_{i=1}^{\infty}\pi_i\sum_{j=1}^{\infty}P_{ij}(d)\left[d + j\eta_Q'(0)\right]. \end{aligned}$$

## Equivalence with Non-Preemptive Priority

$$E\left[\mathcal{W}_{DAP,2}\right] = E\left[\mathcal{W}_{DAP,2}\mathbb{1}\{\mathcal{W}_{DAP,2} \leq d\}\right] + E\left[\mathcal{W}_{DAP,2}\mathbb{1}\{\mathcal{W}_{DAP,2} > d\}\right]$$
$$E\left[\mathcal{W}_{NP,2}\right] = E\left[\mathcal{W}_{NP,2}\mathbb{1}\{\mathcal{W}_{NP,2} \leq d\}\right] + E\left[\mathcal{W}_{NP,2}\mathbb{1}\{\mathcal{W}_{NP,2} > d\}\right]$$

# Class-2 M/M/1 Average Waiting Time

## Removing the Laplace-Stieltjes Transform

$$\begin{aligned} & E\left[\mathcal{W}_{Q,2}\mathbb{1}\{\mathcal{W}_{Q,2} > d\}\right] \\ = & -\frac{d}{ds}E\left[e^{-s\mathcal{W}_{Q,2}}\mathbb{1}\{\mathcal{W}_{Q,2} > d\}\right]\Big|_{s=0} \\ = & \sum_{i=1}^{\infty}\pi_i\sum_{j=1}^{\infty}P_{ij}(d)\left[d + j\eta'_Q(0)\right]. \end{aligned}$$

## Equivalence with Non-Preemptive Priority

$$E\left[\mathcal{W}_{DAP,2}\right] = E\left[\mathcal{W}_{DAP,2}\mathbb{1}\{\mathcal{W}_{DAP,2} \leq d\}\right] + E\left[\mathcal{W}_{DAP,2}\mathbb{1}\{\mathcal{W}_{DAP,2} > d\}\right]$$
$$E\left[\mathcal{W}_{NP,2}\right] = E\left[\mathcal{W}_{NP,2}\mathbb{1}\{\mathcal{W}_{NP,2} \leq d\}\right] + E\left[\mathcal{W}_{NP,2}\mathbb{1}\{\mathcal{W}_{NP,2} > d\}\right]$$

## Average Waiting Time

$$E\left[\mathcal{W}_{NP,2} - \mathcal{W}_{DAP,2}\right] = \sum_{i=1}^{\infty}\pi_i\sum_{j=1}^{\infty}P_{ij}(d)j\ \underbrace{\left[\eta'_{NP}(0) - \eta'_{DAP}(0)\right]}_{\Delta_2}$$

# Class-1 M/M/1 Average Waiting Time

**Result**

$$E\left[\mathcal{W}_{NP,2} - \mathcal{W}_{DAP,2}\right]$$

$$= \Delta_2 \left[ (1-\rho) \sum_{k=0}^{\infty} \frac{e^{-\nu d}(\nu d)^k}{k!} \left( \sum_{j=1}^{k} \gamma_j^{(k)} \right) + \rho e^{-(1-r)(\nu d)} \left( \frac{1}{1-\rho} + r\nu d \right) \right]$$

# Class-1 M/M/1 Average Waiting Time

**Result**

$$E\left[\mathcal{W}_{NP,2} - \mathcal{W}_{DAP,2}\right]$$

$$= \Delta_2 \left[ (1-\rho) \sum_{k=0}^{\infty} \frac{e^{-\nu d}(\nu d)^k}{k!} \left( \sum_{j=1}^{k} \gamma_j^{(k)} \right) + \rho e^{-(1-r)(\nu d)} \left( \frac{1}{1-\rho} + r\nu d \right) \right]$$

**Non-Preemptive Priority**

$$E\left[\mathcal{W}_{NP,2}\right] = \frac{\lambda}{\mu^2(1-\rho_1)(1-\rho)}$$

# Class-1 M/M/1 Average Waiting Time

**Result**

$$E\left[\mathcal{W}_{NP,2} - \mathcal{W}_{DAP,2}\right]$$

$$= \Delta_2 \left[ (1-\rho) \sum_{k=0}^{\infty} \frac{e^{-\nu d}(\nu d)^k}{k!} \left( \sum_{j=1}^{k} \gamma_j^{(k)} \right) + \rho e^{-(1-r)(\nu d)} \left( \frac{1}{1-\rho} + r\nu d \right) \right]$$

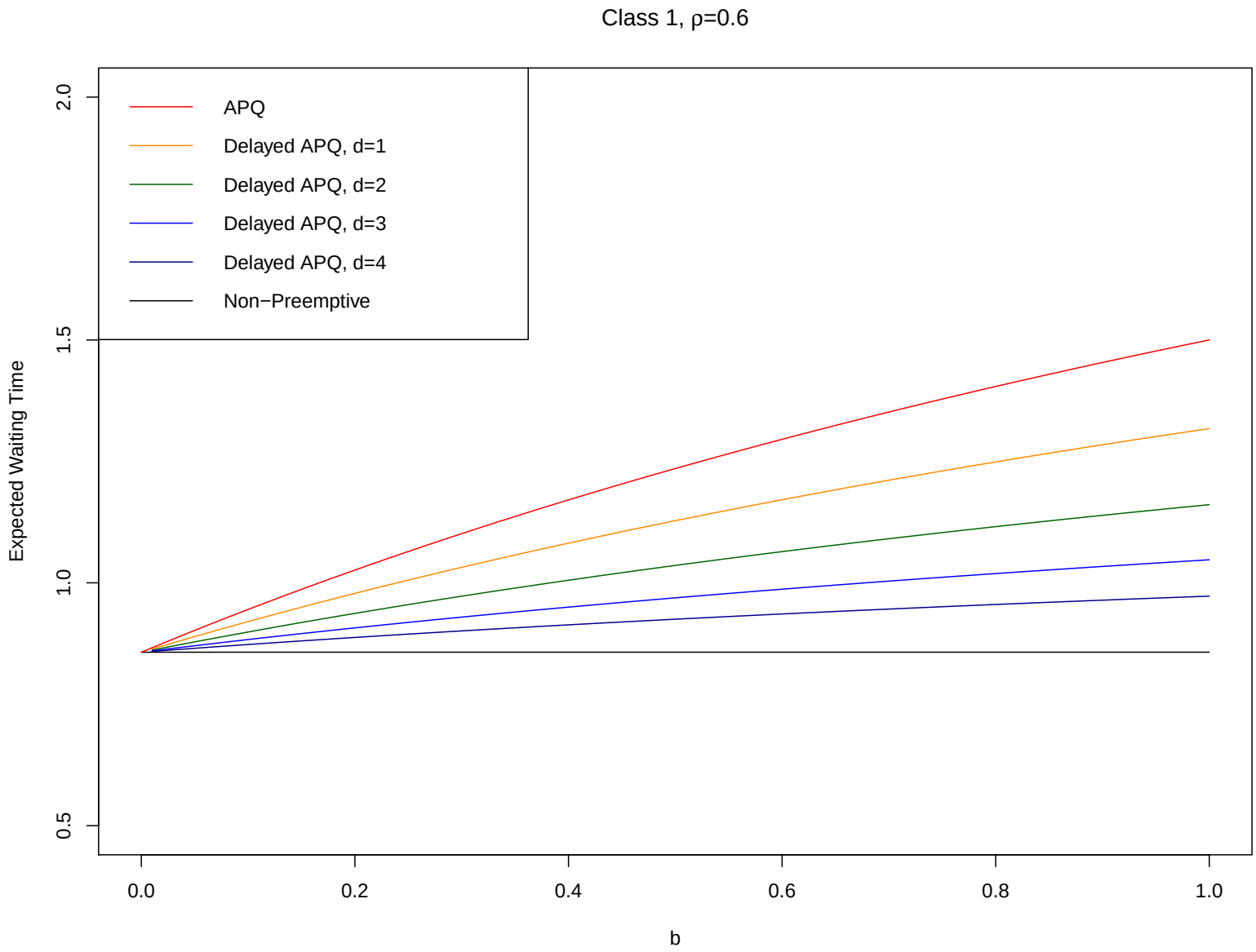
**Non-Preemptive Priority**

$$E\left[\mathcal{W}_{NP,2}\right] = \frac{\lambda}{\mu^2(1-\rho_1)(1-\rho)}$$

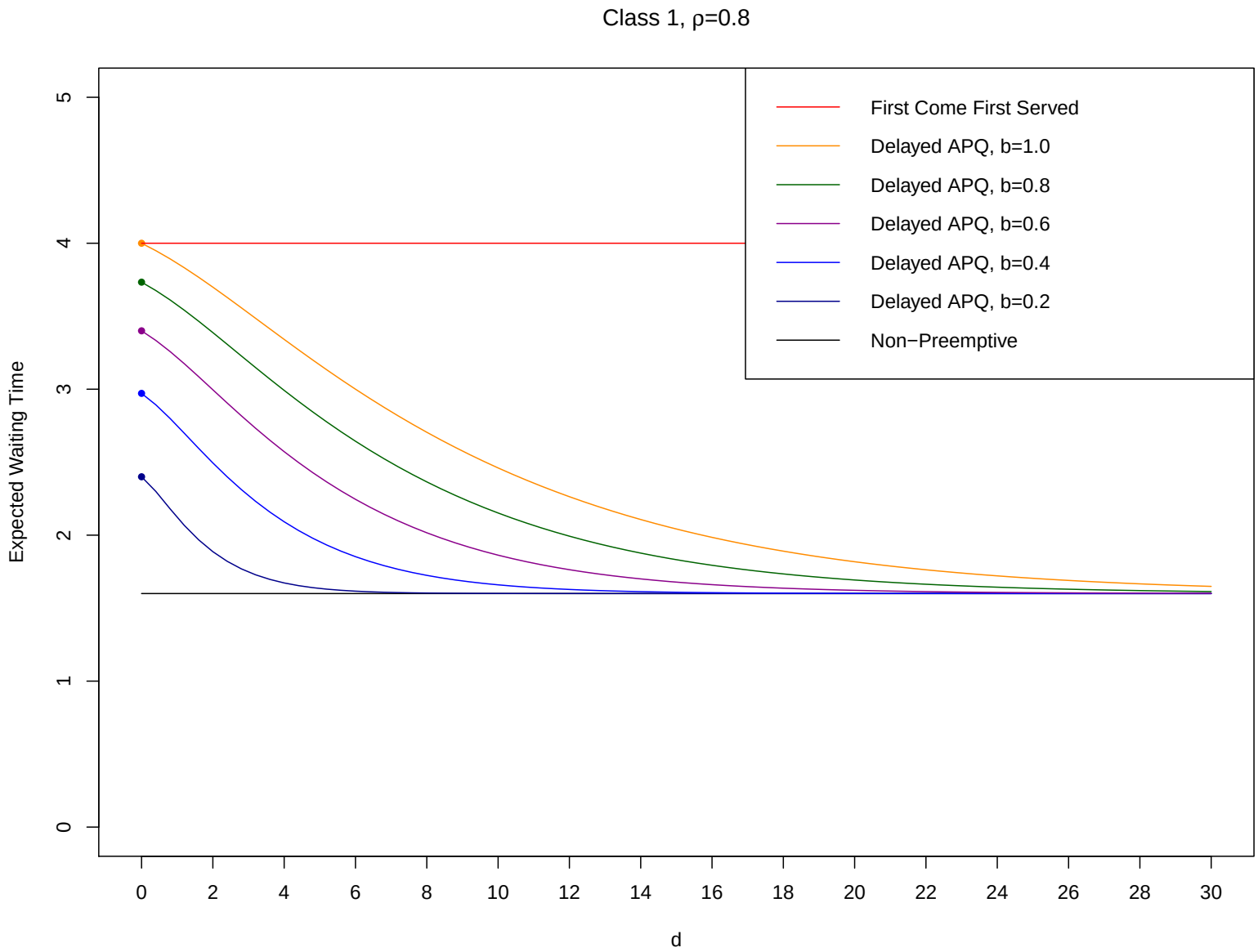
**Conservation Law**

$$\frac{\rho^2}{\mu - \lambda} = \rho_1 E[\mathcal{W}_{DAP,1}] + \rho_2 E[\mathcal{W}_{DAP,2}]$$

# Effect of Accumulation Rate



# Effect of Delay Length



# M/M/c and M/G/1 Extension

## Multiple Servers

- When all servers are busy, the queue is indistinguishable from an M/M/1 delayed APQ with service at rate $c\mu$.
- The probability of all servers being busy is the Erlang-C probability, and can be readily computed.

# M/M/c and M/G/1 Extension

## Multiple Servers

- When all servers are busy, the queue is indistinguishable from an M/M/1 delayed APQ with service at rate $c\mu$.
- The probability of all servers being busy is the Erlang-C probability, and can be readily computed.

## General Service

- Without the assumption of exponential service, the accreditation interval length $\eta(s)$ may be unknown.
- A special case is deterministic service, where all customers have a service time of exactly $1/\mu$.
- The residual service time now will have a distribution which depends on $N_d$, since if there are more customers it implies the service has been going on for longer. This is tractable to compute, although it remains to efficiently implement an algorithm to do so.

# Summary

- ★ The Delayed APQ allows Class-1 customers to benefit more than the APQ while not being as harsh to Class-2 customers as the Non-Preemptive Priority Queue.
- ★ The Delayed APQ Class-2 waiting time is equivalent to the Non-Preemptive waiting time prior to time $d$.
- ★ After time $d$, the savings in the Delayed APQ for Class-2 can be completely explained by the shorter accreditation interval length.
- ★ The Delayed APQ Class-1 average waiting time can be calculated using the conservation law without understanding how the process develops after time $d$.