Required work in the M/M/1 queue

Samuli Aalto
Helsinki University of Technology
Finland

Contents

- Motivation: Required work. Where do we need it?
- Main result: Equilibrium distribution.
- Proof: Reversibility revisited.

Definitions (the FIFO case)

- **Unfinished work**, U (in time units)
  - service times of the **waiting** customers **plus**
  - **residual** service time of the **served** customer
- **Finished work**, V (in time units)
  - **elapsed** service time of the **served** customer
- **Required work**, $Z = U + V$ (in time units)
  - service times of **all** the cusomers in the system

Motivating problem

- How to allocate (and release) memory for **variable length** packets e.g. in the output buffer of a IP router?
- The following three allocation schemes are considered:
  - static
  - fully dynamic
  - dynamic

Static allocation scheme

- **allocate** a memory block of **maximum** (i.e. fixed) length for a packet when it arrives
- **release** the block **as a whole** as soon as the packet has been transmitted (totally)
- light processing but wasteful memory usage
- in queueing terms, the interesing variable is
  - the number of customers in the system, N

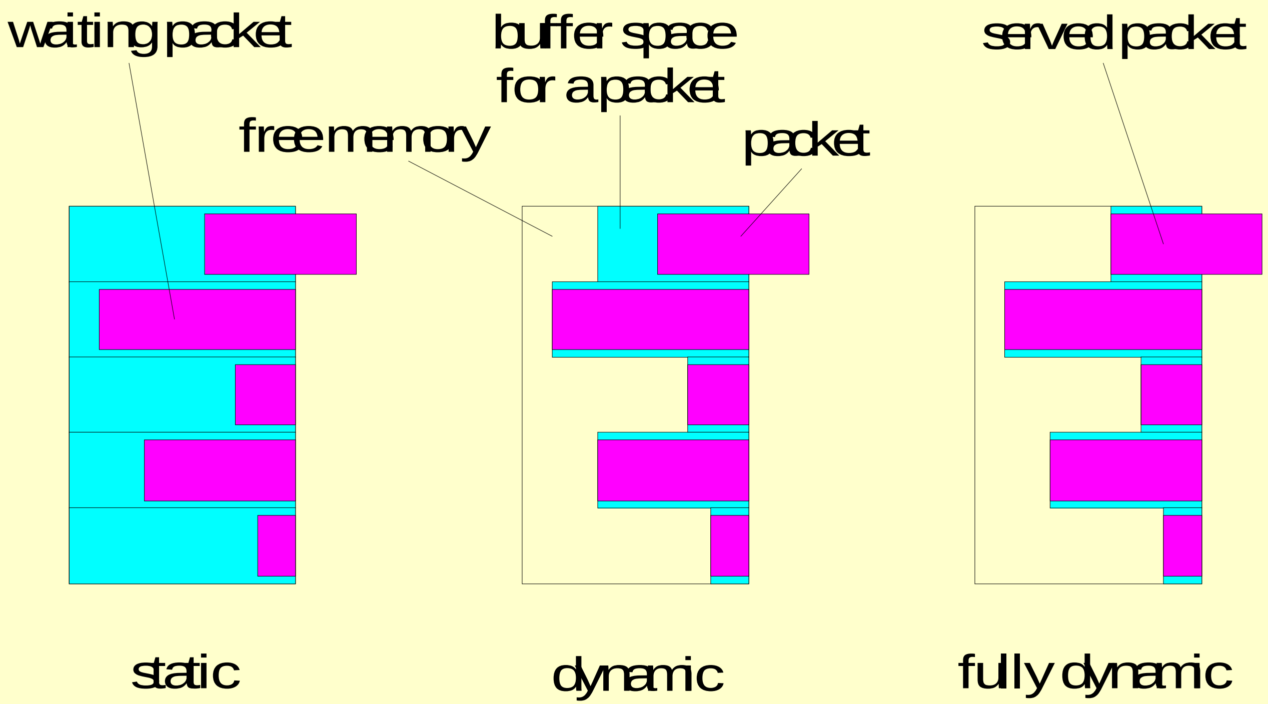
Fully dynamic allocation scheme

- **allocate** a memory block of **actual** (i.e. variable) length for a packet when it arrives
- **release** the block **gradually** as the transmission of the packet proceeds
- efficient memory usage but heavy processing
- in queueing terms, the interesting variable is
  - the amount of unfinished work, U

Dynamic allocation scheme

- **allocate** a memory block of **actual** (i.e. variable) length for a packet when it arrives
- **release** the block **as a whole** as soon as the packet has been transmitted (totally)
- a compromise between the former two:
  not so heavy processing and still reasonably efficient memory usage
- in queueing terms, the interesting variable is
  - the sum of the unfinished and finished work, $U+V$
- this is called the **required work**, Z

Three memory allocation principles



Contents

- Motivation: Required work. Where do we need it?
- Main result: Equilibrium distribution.
- Proof: Reversibility revisited.

M/M/1 basics

- Assume
  - stable M/M/1 queue with FIFO queueing discipline
- Unfinished work, U

$$E[\exp(-sU)] = (1-\rho) + \rho\left(\frac{\mu(1-\rho)}{\mu(1-\rho)+s}\right)$$

- Finished work, V

$$E[\exp(-sU)] = (1-\rho) + \rho\left(\frac{\mu}{\mu+s}\right)$$

Main result (1)

- Required work, $Z = U + V$

$$E[\exp(-sZ)] =$$

$$(1-\rho) + \rho\left(\frac{\mu(1-\rho)}{\mu(1-\rho)+s}\right)\left(\alpha_1 \frac{\mu\theta_1}{\mu\theta_1 + s} + \alpha_2 \frac{\mu\theta_2}{\mu\theta 2 + s}\right) \quad (1)$$

- Note that U and V are not independent

Main result (2)

$$\alpha_1 = \frac{\sqrt{4+\rho} + \sqrt{\rho}}{(2+\rho)\sqrt{4+\rho} + (4+\rho)\sqrt{\rho}}$$

$$\alpha_2 = \frac{(1+\rho)\sqrt{4+\rho} + (3+\rho)\sqrt{\rho}}{(2+\rho)\sqrt{4+\rho} + (4+\rho)\sqrt{\rho}} = 1-\alpha_1$$

$$\theta_1 = \frac{1}{2}\left(2+\rho-\sqrt{(4+\rho)\rho}\right)$$

$$\theta_2 = \frac{1}{2}\left(2+\rho+\sqrt{(4+\rho)\rho}\right) = \theta_1^{-1}$$

Interpretation

- Here $\alpha_1 + \alpha_2 = 1$. Thus

$$Z = I(\zeta_0 + J\zeta_1 + (1 - J)\zeta_2)$$

where I, J, ζ_0, ζ_1, ζ_2 are independent with

$$I \sim \text{Bernoulli}(\rho)$$
$$J \sim \text{Bernoulli}(\alpha_1)$$

$$\zeta_0 \sim \text{Exp}(\mu(1 - \rho))$$
$$\zeta_1 \sim \text{Exp}(\mu\theta_1)$$
$$\zeta_2 \sim \text{Exp}(\mu\theta_2)$$

- In addition, $\theta_1\theta_2 = 1$ and $\alpha_1/\theta_1 + \alpha_2/\theta_2 = 1$.

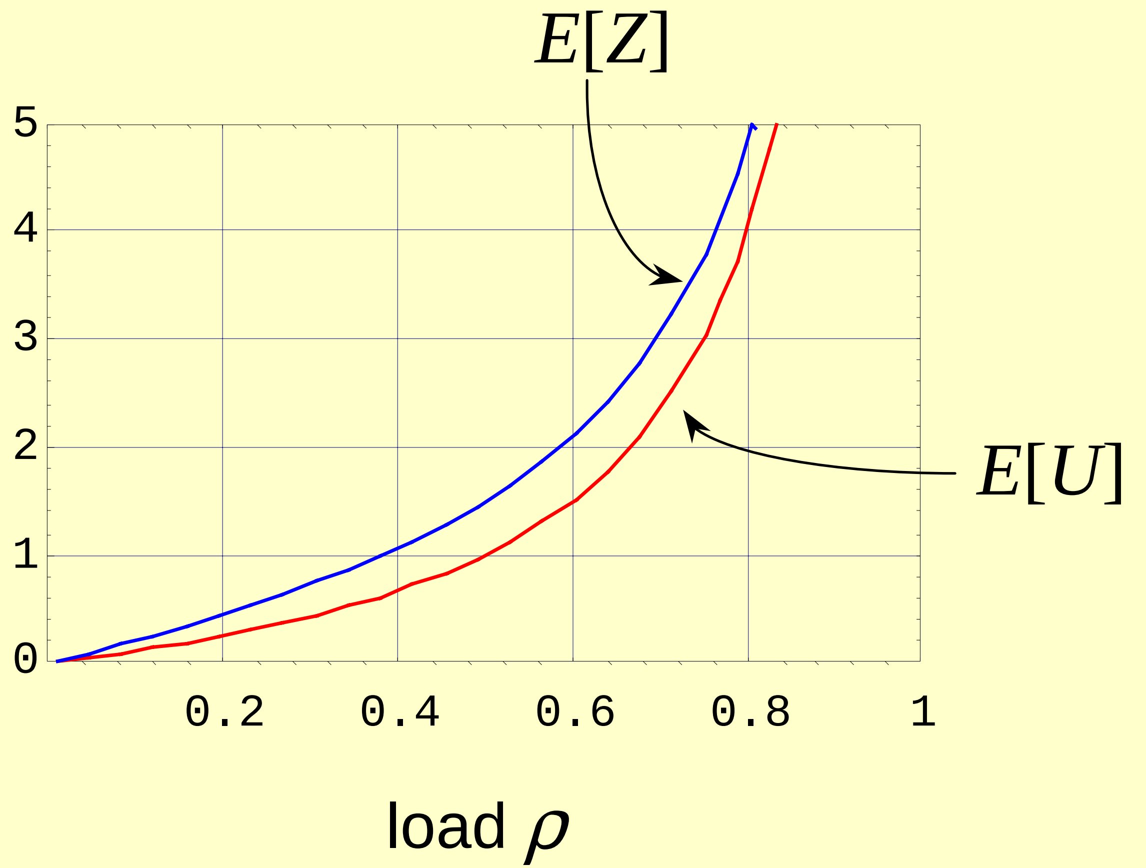
Corollary

$$P\{Z > z\} = \rho e^{-\mu(1-\rho)z} + \sum_{i=1}^{2} \frac{\rho(1-\rho)\alpha_i}{1-\rho-\theta_i}(e^{-\mu\theta_i z} - e^{-\mu(1-\rho)z})$$

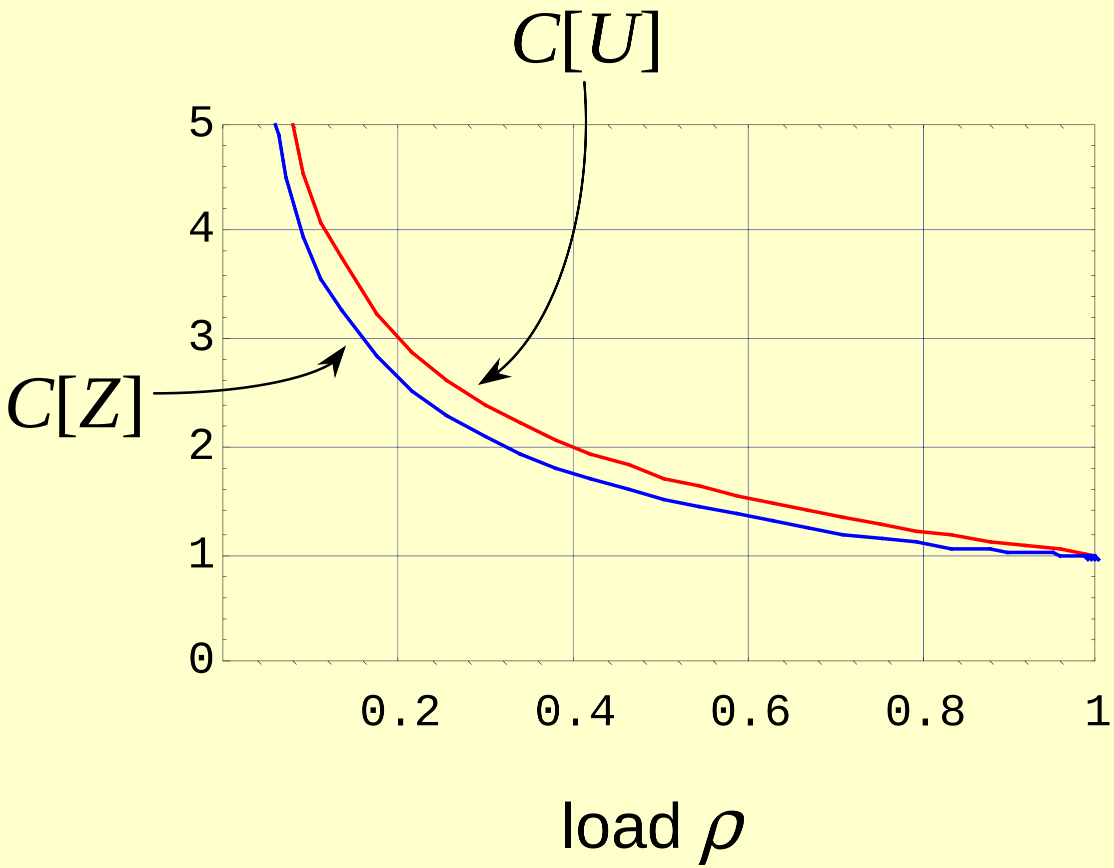
$$E[Z] = \frac{\rho(2-\rho)}{\mu(1-\rho)}$$

$$D^2[Z] = \left(\frac{\rho(2-\rho)}{\mu(1-\rho)}\right)^2 \frac{6-8\rho+2\rho^2+\rho^3}{\rho(2-\rho)^2}$$

Expectation



Coefficient of variation



Contents

Proof (1)

- Essential difficulty in deriving the distribution of Z:
  - dependence between N and V
- Let

$$F_n(v) = P\{N = n, V \le v\}$$

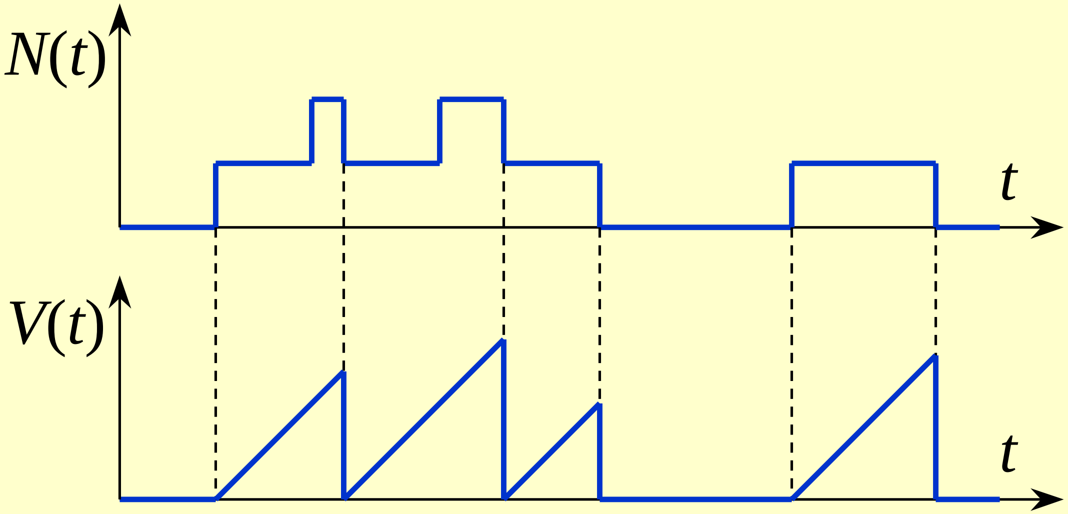
Proof (2)

- Given N and V, the required work Z is the sum of the following **conditionally independent** components
  - N-1 service times of the waiting customers
  - residual service time of the served customer
  - elapsed service time V of the served customer
- Thus,

$$E[\exp(-sZ)] = (1-\rho) + \sum_{n=1}^{\infty} \int_0^{\infty} F_n(dv) e^{-sv} \left(\frac{\mu}{\mu + s} \right)^n \quad (2)$$

Proof (3)

- $(N(t), V(t))$ constitutes a Markov process



Proof (4)

- Balance equations

$$F_n'(v) = \lambda F_{n-1}(v) - (\lambda + \mu) F_n(v) + \mu F_{n+1}(\infty) \qquad (3)$$

- Boundary values

$$F_0(v) = 1 - \rho; F_n(0) = 0, F_n(\infty) = (1-\rho)\rho^n, n = 1,2,\ldots$$

- Solution

$$F_n(v) = (1-\rho)\rho^n \left(1 - \sum_{k=0}^{n-1} \frac{(\mu v)^k}{k!} e^{-(\lambda+\mu)v}\right)$$

- Apply this to (2) to get the Laplace transform (1).

Conditional distribution

- It follows that

$$P\{V \leq v | N = n\} = 1 - \sum_{k=0}^{n-1} \frac{(\mu v)^k}{k!} e^{-(\lambda+\mu)v} \qquad (4)$$

- Thus,

$$V | N \sim \min\{X_1 + \ldots + X_N, Y\}$$

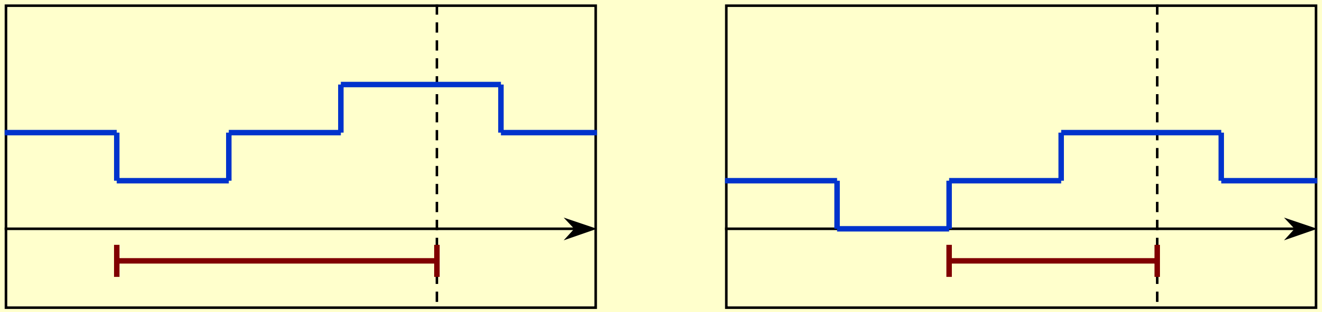
where all $X_i \sim \mathrm{Exp}(\mu)$ and $Y \sim \mathrm{Exp}(\lambda)$ are independent

- This can be explained by a reversibility argument

Reversibility argument

- The elapsed service time in the original process corresponds to the time until
  - a new customer arrives (Y) or
  - the queue becomes empty ($X_1 + \ldots + X_N$)

  (whichever occurs first) in the reversed process



Summary

- The M/M/1 queue with the FIFO queueing discipline was considered.
- Introduction of a new variable, the required work, was motivated by a dynamic memory allocation scheme.
- Equilibrium distribution of the required work was derived.
- Reversibility argument was utilized.

LOPPU