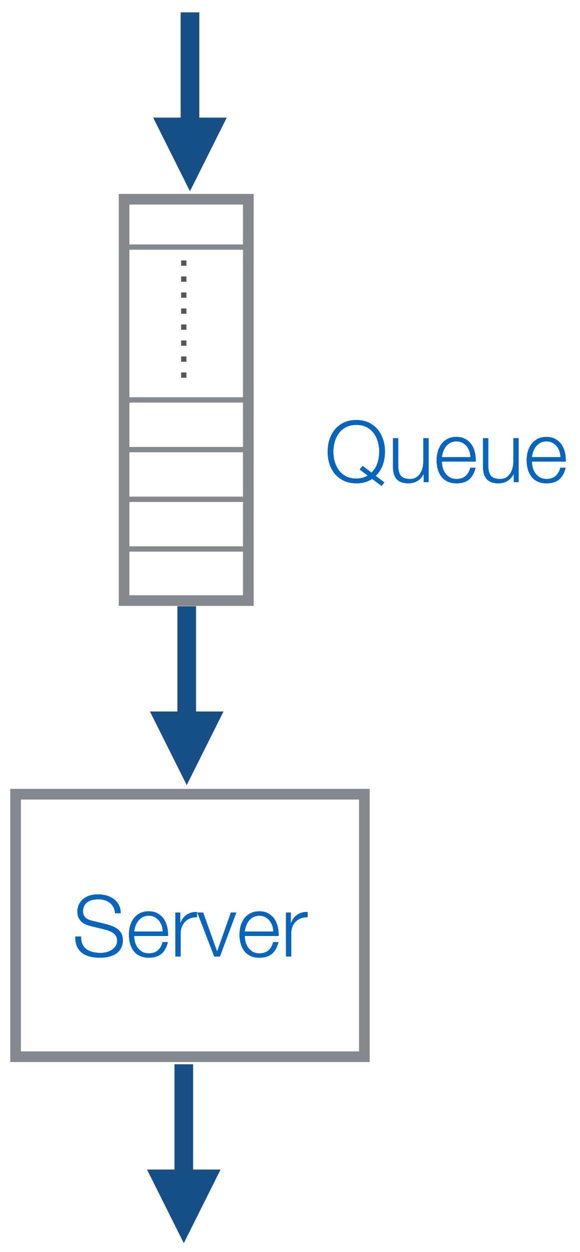
CS 925
Lecture 4
Traffic Management

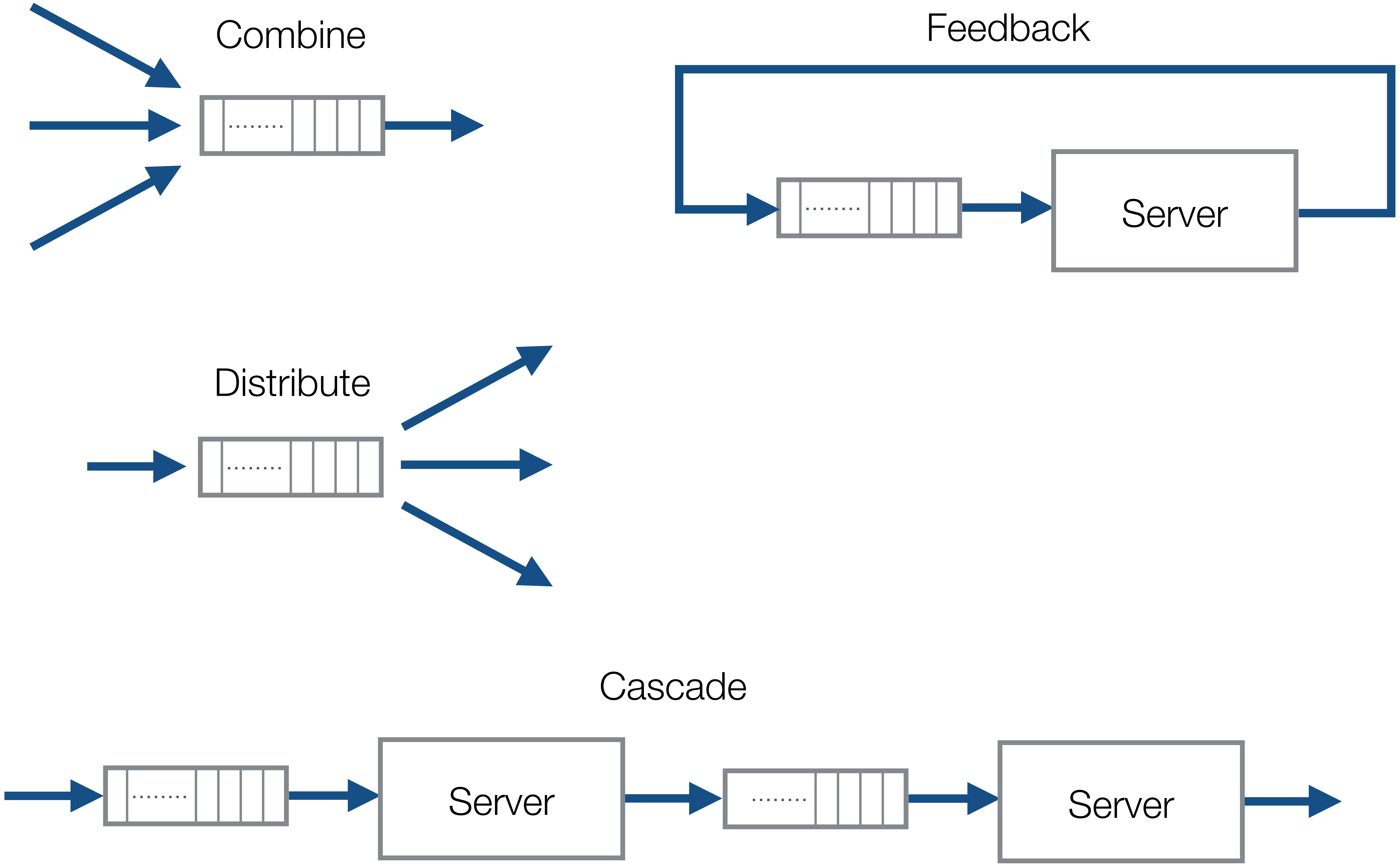
Thursday, February 1, 2024

Queuing System

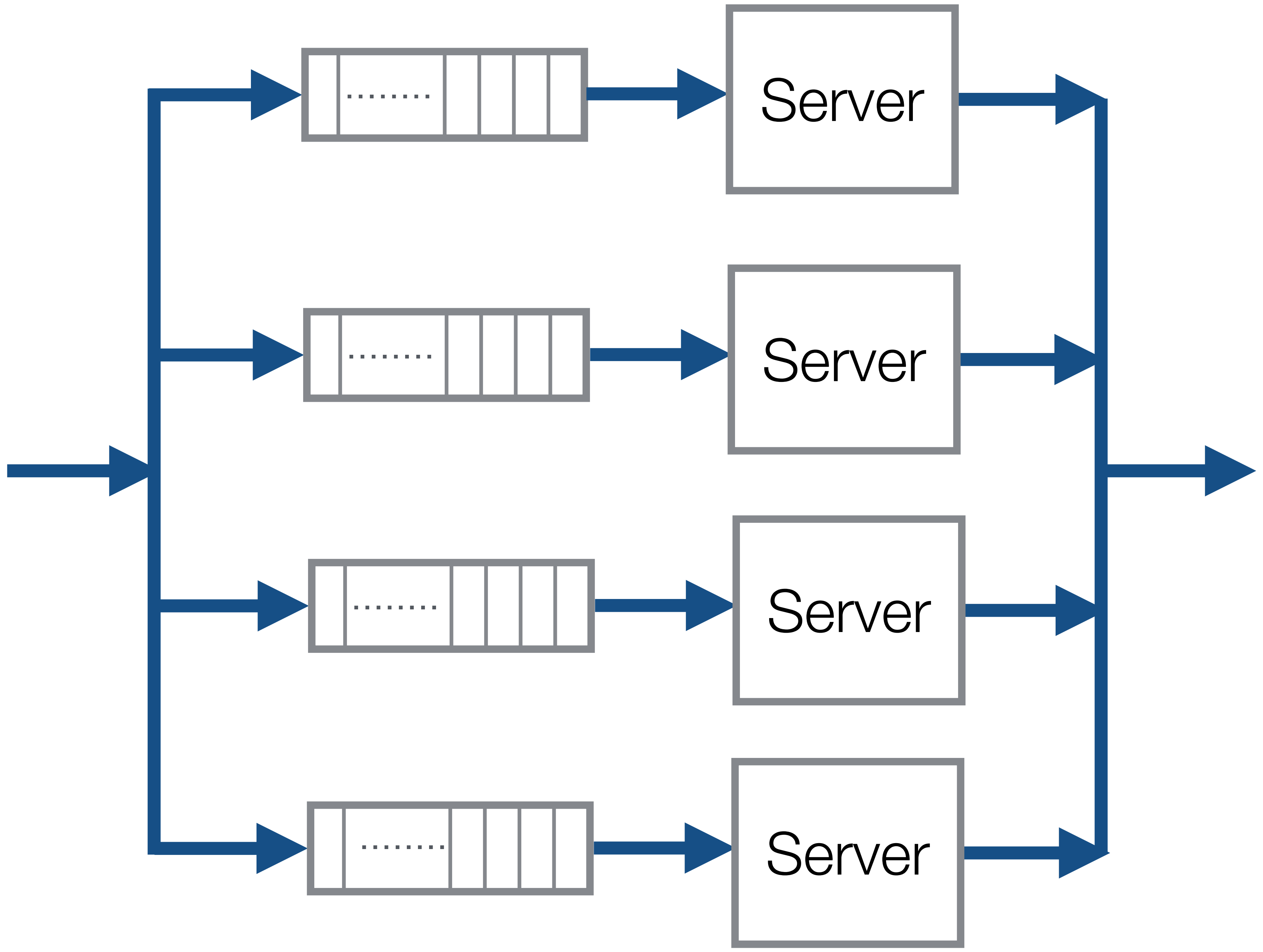
- A model of a system where entities wait (in a queue) for a service or a resource (server)
- Entities
  - packets, messages, tasks, people, …
- Service / Resource
  - switching fabric of a networking device
  - transmission line
  - protocol stack
  - …



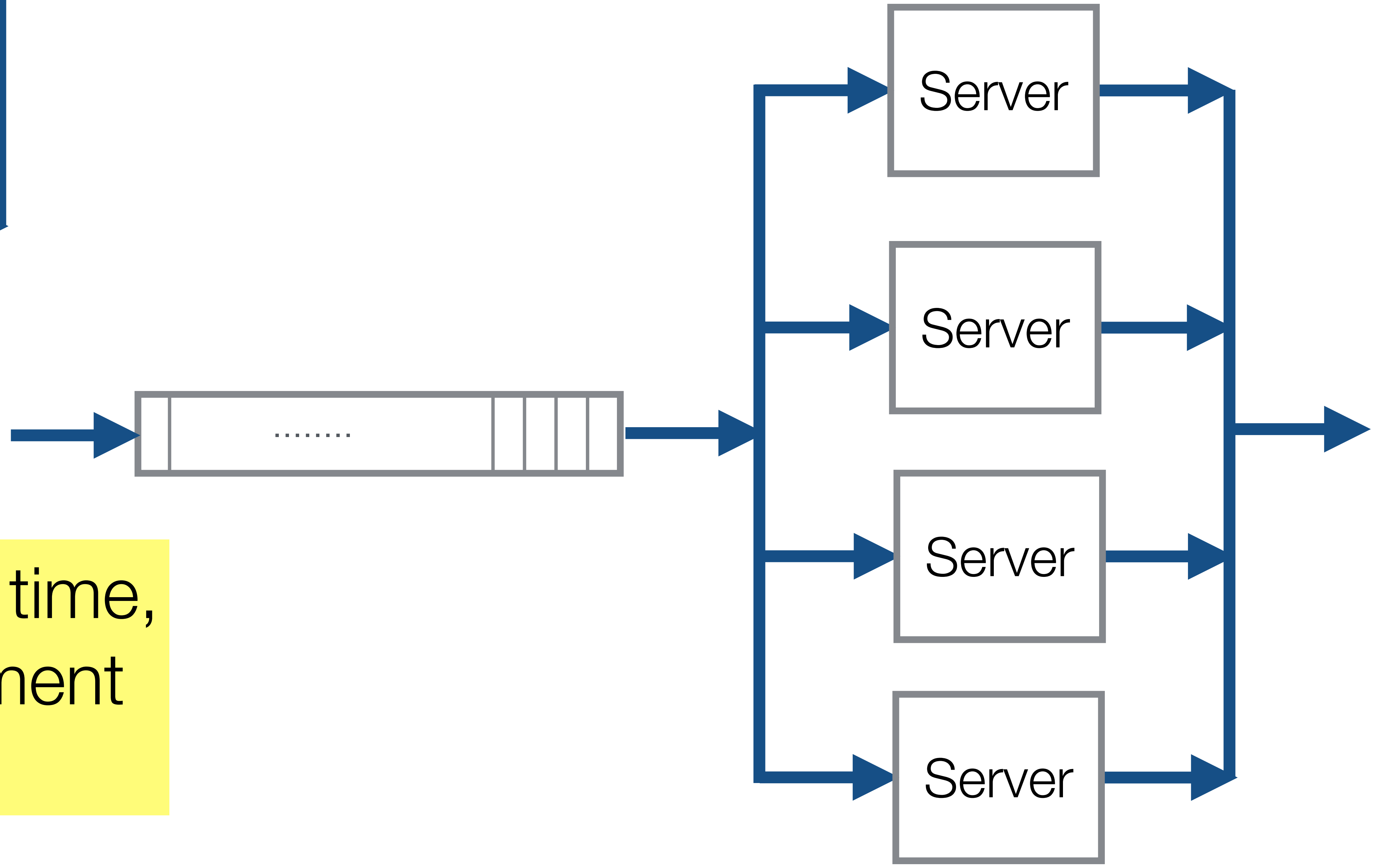
Queuing System



Exercise



Assume that all servers are the same

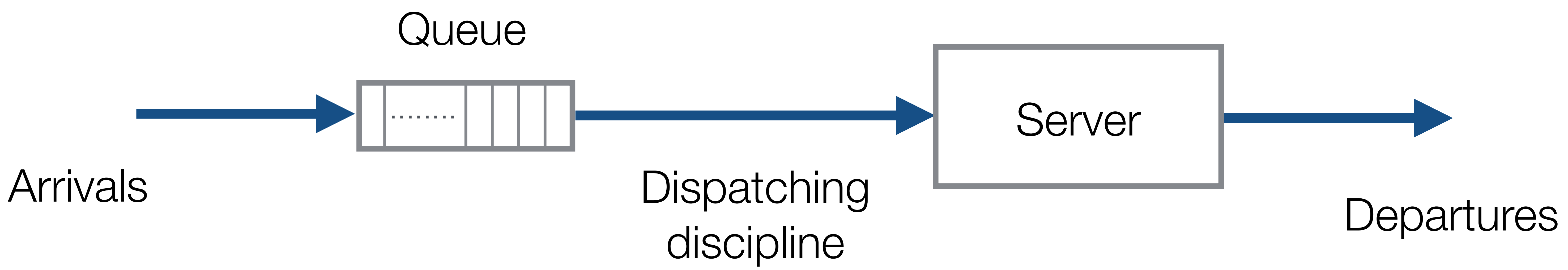


Considering expected wait time, which one of the arrangement is “better”?

Queueing System

- Challenges:
  - Randomly arriving “requests”
  - Random “service time”
  - Random number of requests
  - ….
- Goal: performance evaluation
  - Average performance
  - Break down point

Single Server Queue



- Model characteristics:
  - Arrival rate λ
  - Number of waiting items w, queue waiting time T_w
  - Server utilization ρ, service time T_s, service rate μ
  - Number of items in the system r, residency time T_r
  - Item population, queue size, dispatching discipline, probability distributions, …

Kendall Notation

- X / Y / N
  - X - distribution of inter-arrival times
  - Y - distribution of service time
  - N - number of servers
- Values for X and Y
  - G - general
  - M - exponential (memoryless)
  - D - deterministic

Example:
M/D/1 - a single server queuing system with exponential arrivals and deterministic service time

Goals of Queuing Analysis

- With some simplification…
- Given:
  - Arrival rate λ
  - Service time T_s or service rate μ
- Find:
  - number of items in the queue w or the system r
  - waiting time in the queue T_w or the system T_r
- Mean values, std deviations, etc.

Basic Queuing Relationships

- Service time and service rate:

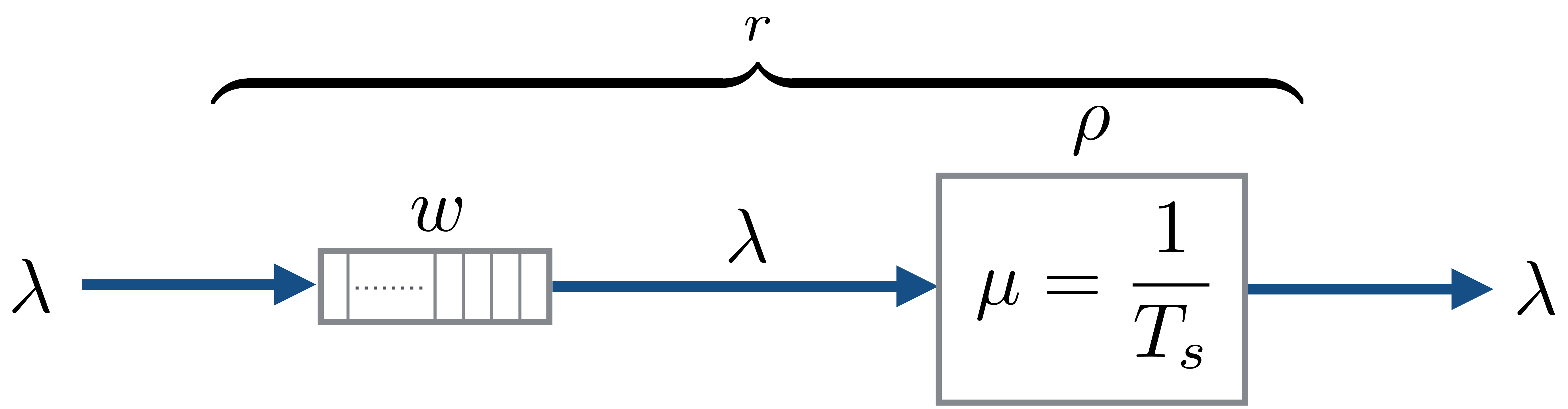
$$\mu = \frac{1}{T_s}$$

- Server utilization:
  - percentage of time the server is in use

$$\rho = \frac{\lambda}{\mu} = \lambda T_s$$

- Number of items in the system:

$$r = w + \rho$$

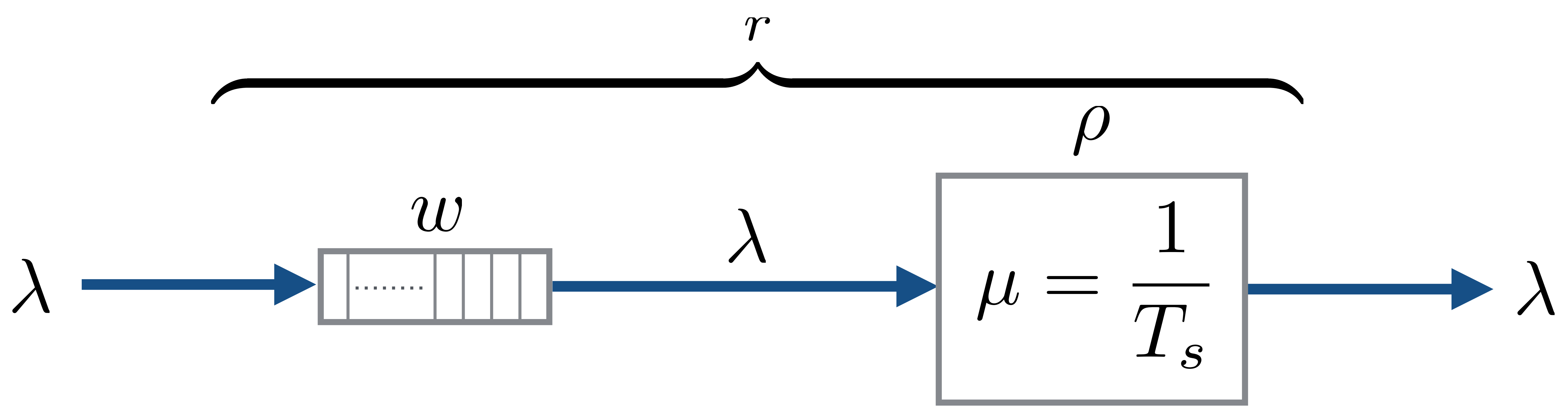


Basic Queuing Relationships

- Little's formula: $w = \lambda T_w$

$$r = w + \rho = \lambda T_w + \frac{\lambda}{\mu} = \lambda T_w + \lambda T_s = \lambda(T_w + T_s) = \lambda T r$$

- The mean number of items in a queuing system depends only on the mean arrival rate and the mean waiting time



M/M/1 Queue

- Given:
  - exponentially distributed inter-arrival time with mean $1/\lambda$
  - exponentially distributed service time with mean $1/\mu = T_s$
  - the system is stable $\lambda < \mu$
- Find the mean number of items on the system/queue and the mean waiting time in the system/queue

$$r = ? \qquad T_r = ? \qquad w = ? \qquad T_w = ?$$

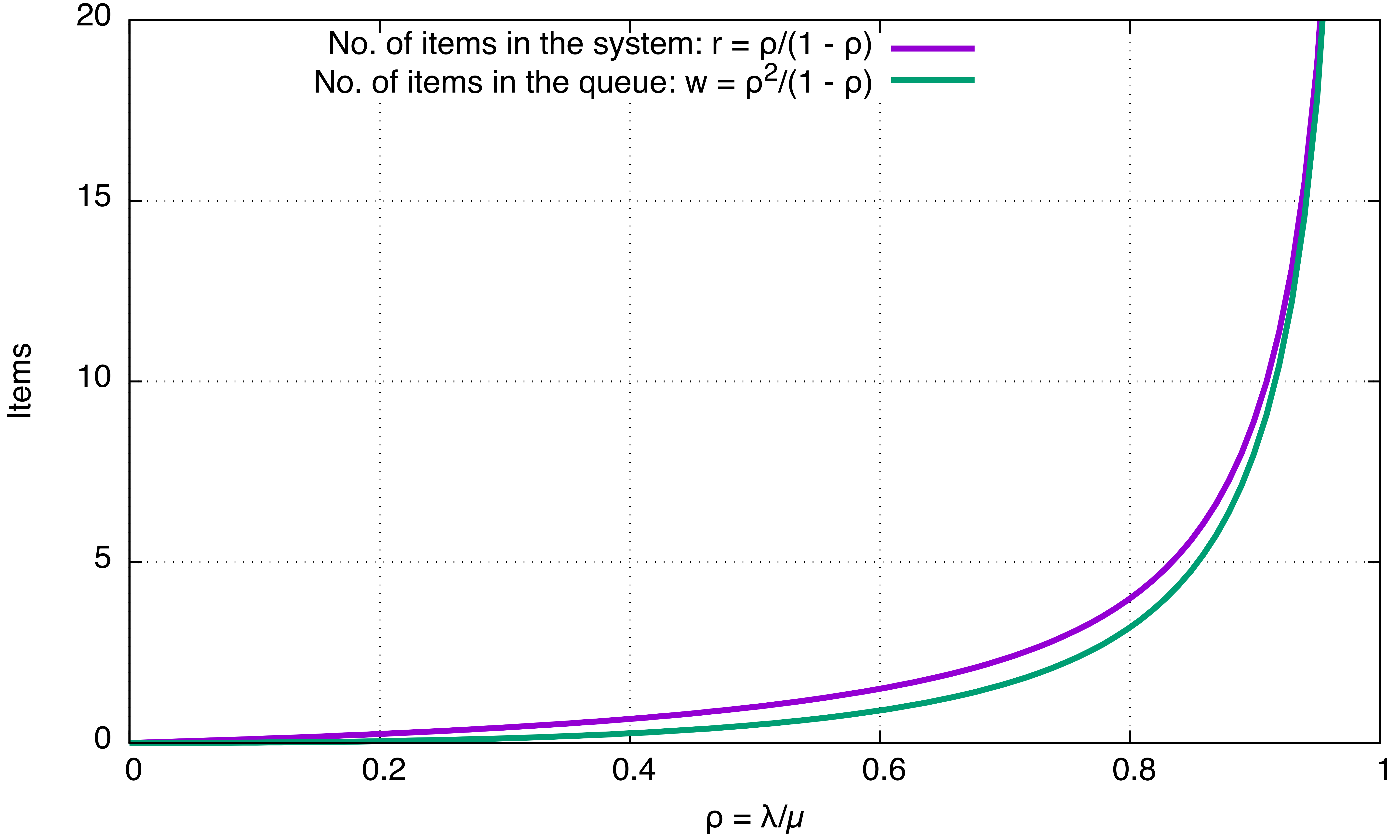
M/M/1 Queue

- Summary of results:

$$r = \frac{\rho}{1-\rho} \qquad w = \frac{\rho^2}{1-\rho}$$

$$T_r = \frac{1}{\mu-\lambda} \qquad T_w = \frac{\lambda}{\mu(\mu-\lambda)}$$

M/M/1 Queue



$$r = \frac{\rho}{1-\rho} \qquad w = \frac{\rho^2}{1-\rho}$$