CSC 2232: Topics in Computer System Performance and Reliability

Bianca Schroeder

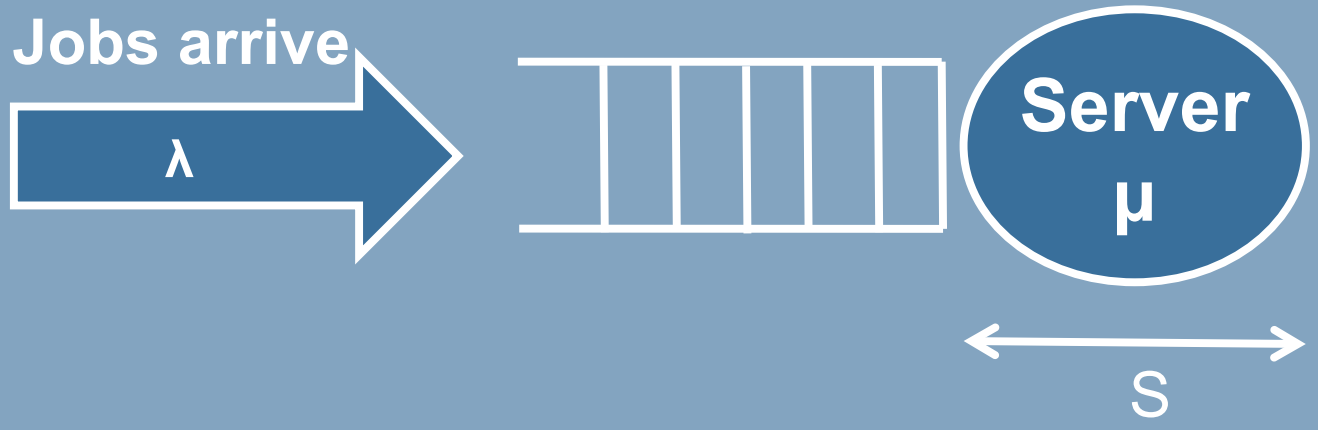
Department of Computer Science
University of Toronto

TODAY'S AGENDA

- More operational laws
- Modification analysis

- After this lecture, you can don a suit and call yourself a systems consultant!

THE SINGLE SERVER NETWORK



- Average arrival rate: λ = 1 job / 6 sec
- Mean interarrival time: $1/\lambda$ = 6 sec
- Service Requirement S
- Mean service time: E[S] = 3 sec
- Average service rate: $\mu = 1/E[S]$ = 1 job / 3 sec
- Service Order: FCFS (First-Come-First-Serve)

PERFORMANCE METRICS

- Response time: T_S
  - Turnaround time, flow time, time in system
  - Main interest: E[T], also Var[T]
- Waiting time: T_Q
  - Time in queue
- Number of jobs in system: N_S
- Number of jobs in queue: N_Q

PERFORMANCE METRICS

Metrics:
T_S
N_S
T_Q
N_Q

- Definitions:

- U_i: Fraction of time device i is busy
- X_i: The rate of completions at device i (in jobs/ sec)

- How does X_i relate to U_i?

HOW DOES X_I RELATE TO U_I

X_i = Mean rate of completion

= E{Rate of comp
P{server busy}

+ E{Rate of comp
P{server is idle}

= μ * U_i

How does this change with changes in
- **Job size distribution**
- **Interarrival time distribution**
- **Service order**

$$X_i = \mu * U_i$$

THE UTILIZATION LAW

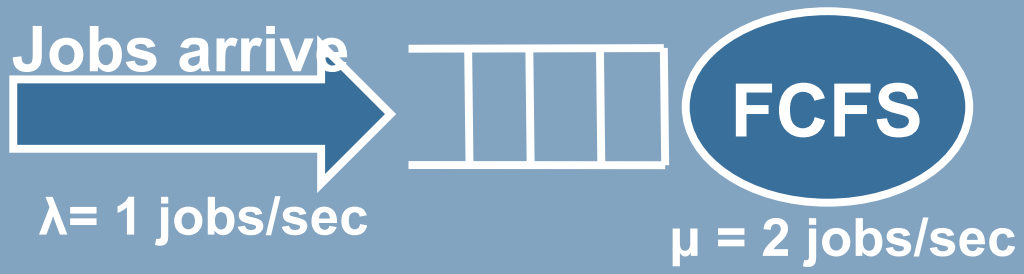
$$X_i = \mu * U_i$$

or equivalently

$$U_i = X_i * E[S]$$

WHAT IS THROUGHPUT?

$$X_i = \mu * U_i$$



- What is U?
  - Fraction of time server is bu
  = mean service time / mean tim etw. Arrivals
  = (1/μ)/(1/λ) = λ/μ

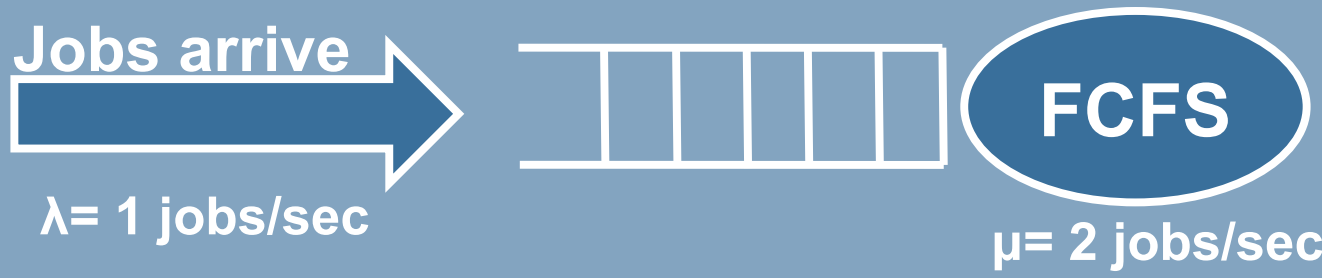
Throughput does not depend on service rate!

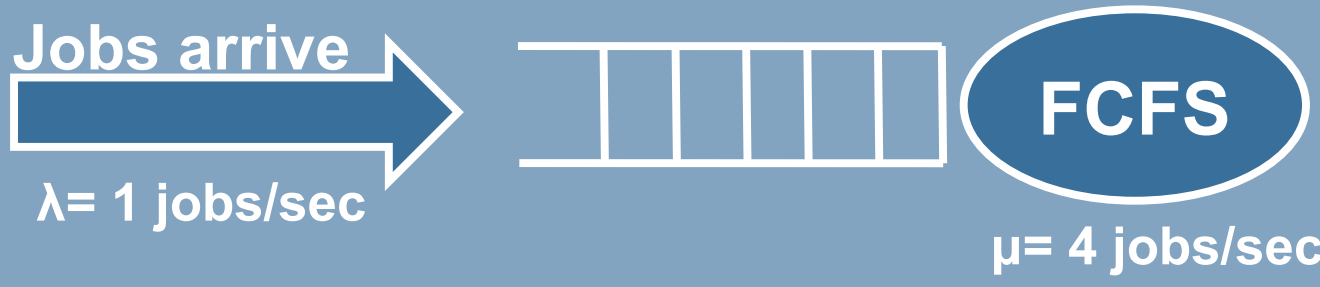
$$X_i = \lambda$$

BACK TO OUR OLD EXAMPLE!

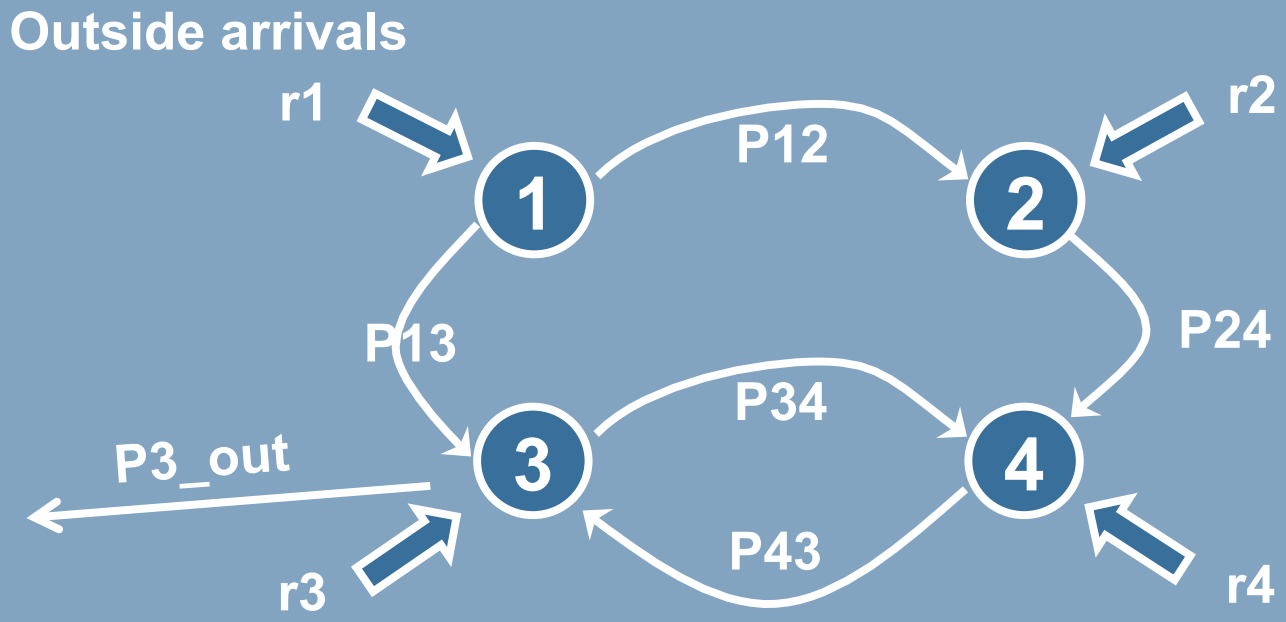
Metrics:
T_S
N_S
T_Q
N_Q
X
U

$$X_i = \mu * U_i = \lambda$$



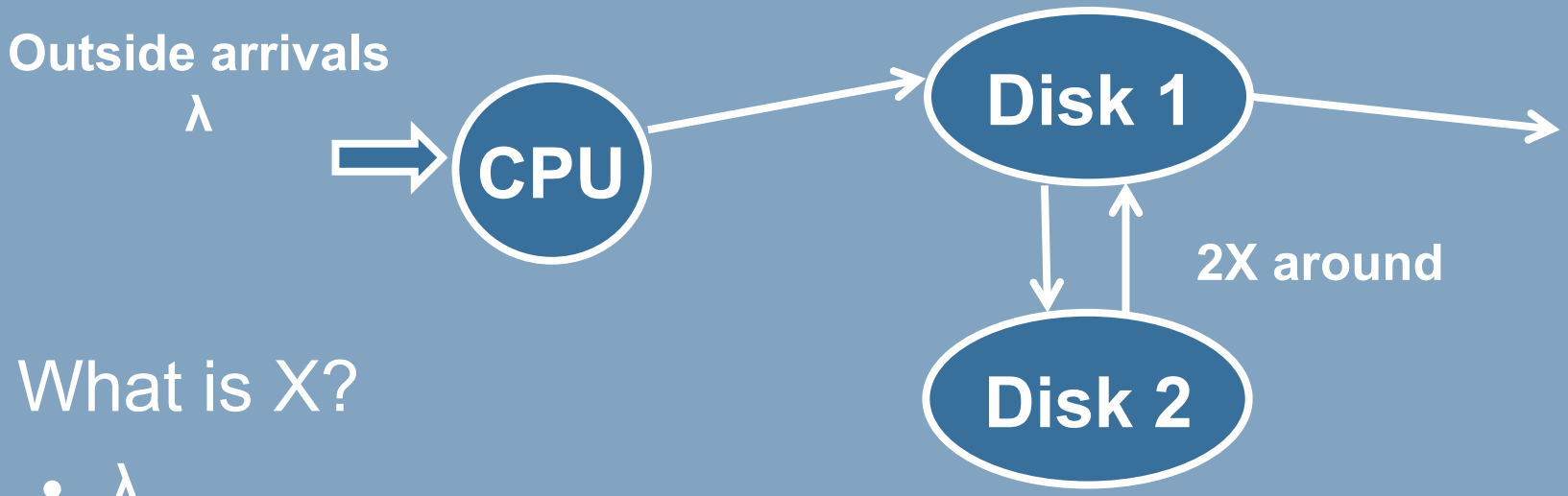


EXAMPLE 2: Network-of-queues with probabilistic routing



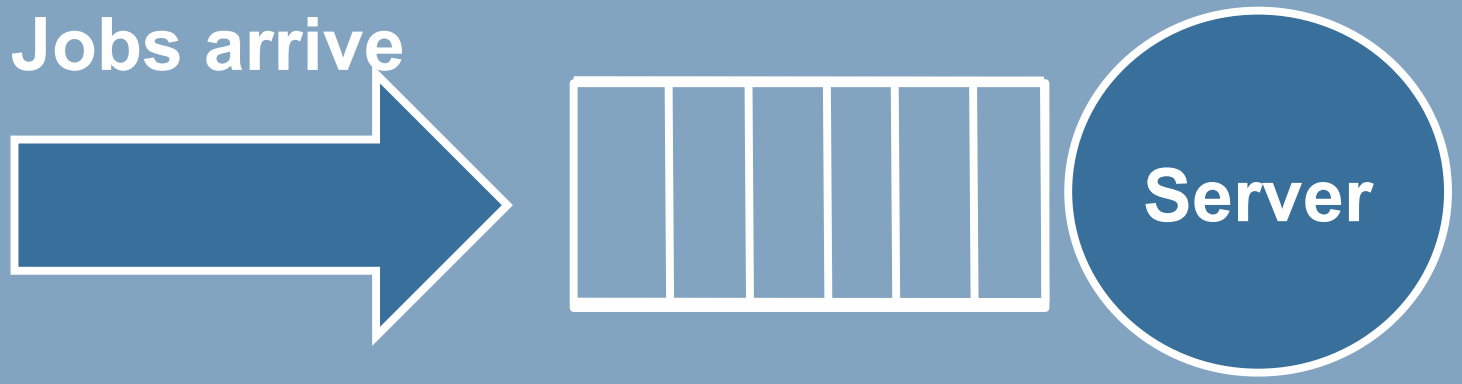
- What is X?
  - ∑ r_i
- What is X_i?
  - X_i = λ_i = r_i + ∑ λ_j * Pji

EXAMPLE 3: Network-of-queues with non-probabilistic routing



- What is X?
  - λ
- What is X_disk1
  - 3λ
- What is X_disk2
  - 2λ

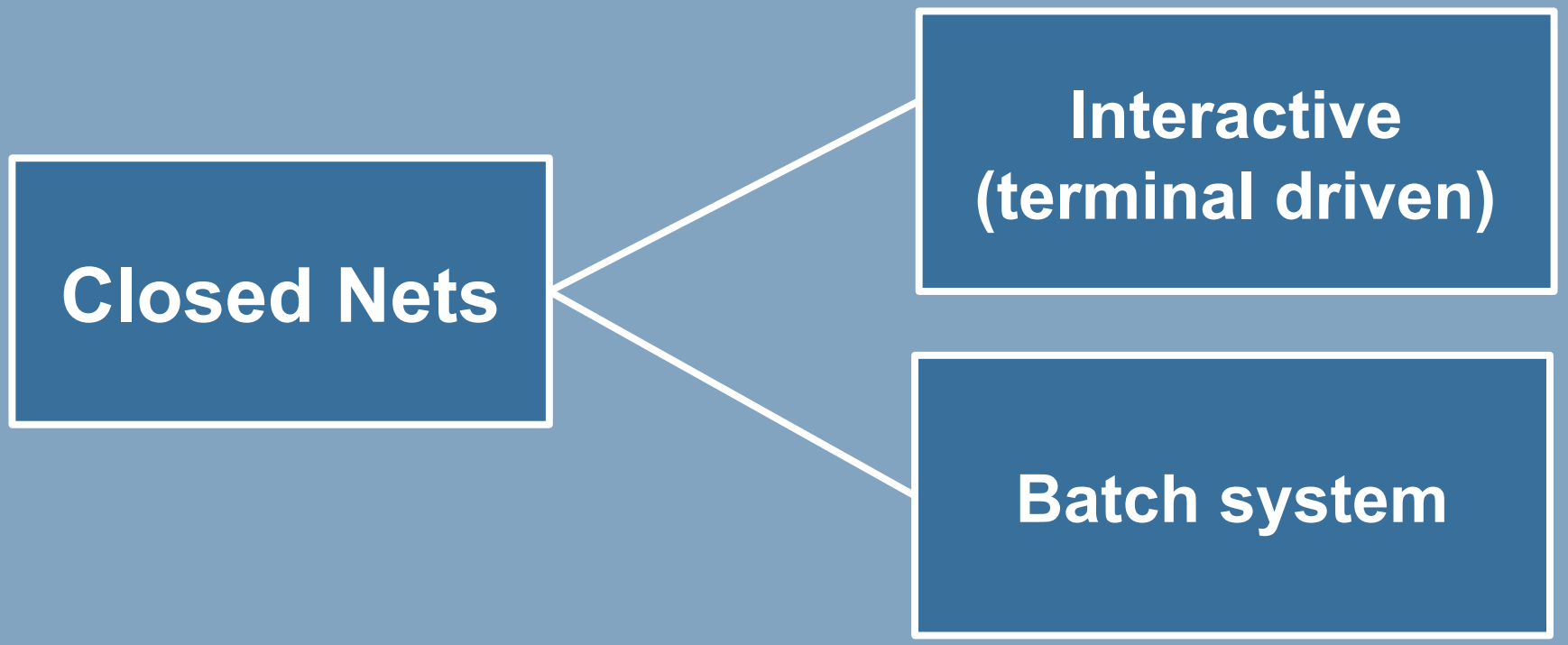
EXAMPLE 4: Finite buffer



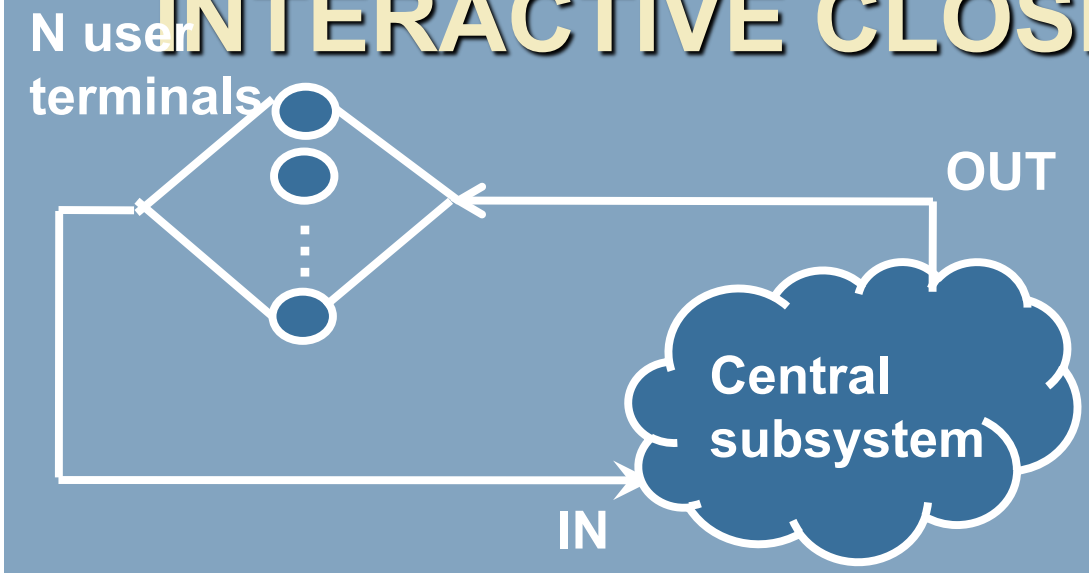
- Space in queue is limited to n jobs
- What is X?
  - $X = U^*\mu$
- But U is no longer λ/μ
- Need Markov chains …

CLOSED SYSTEMS

- Closed networks have no external arrivals or departures



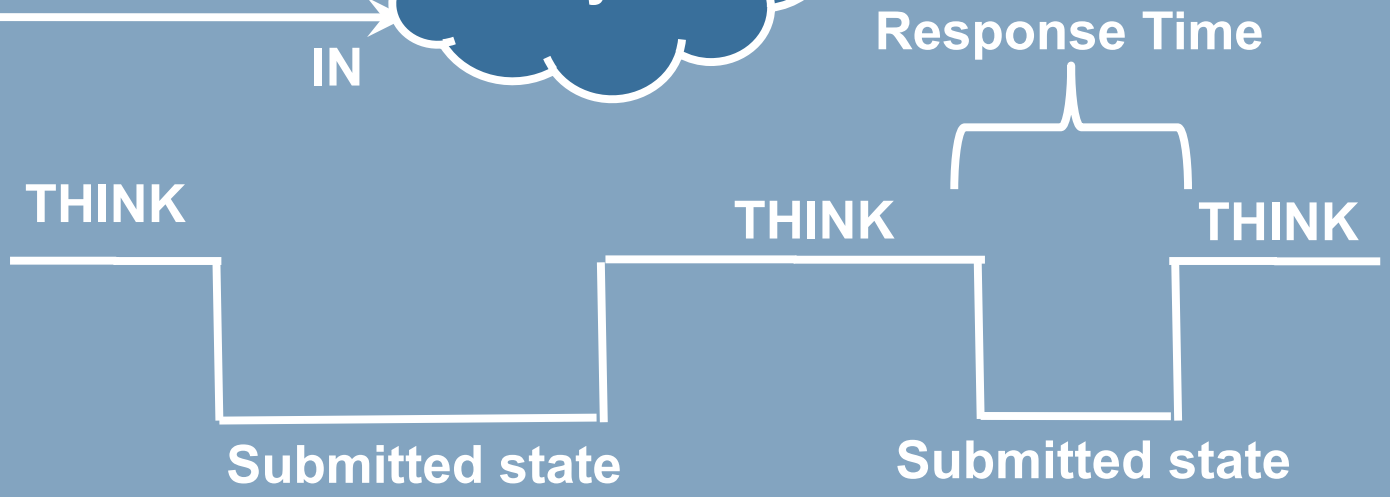
INTERACTIVE CLOSED SYSTEMS



Parameters:

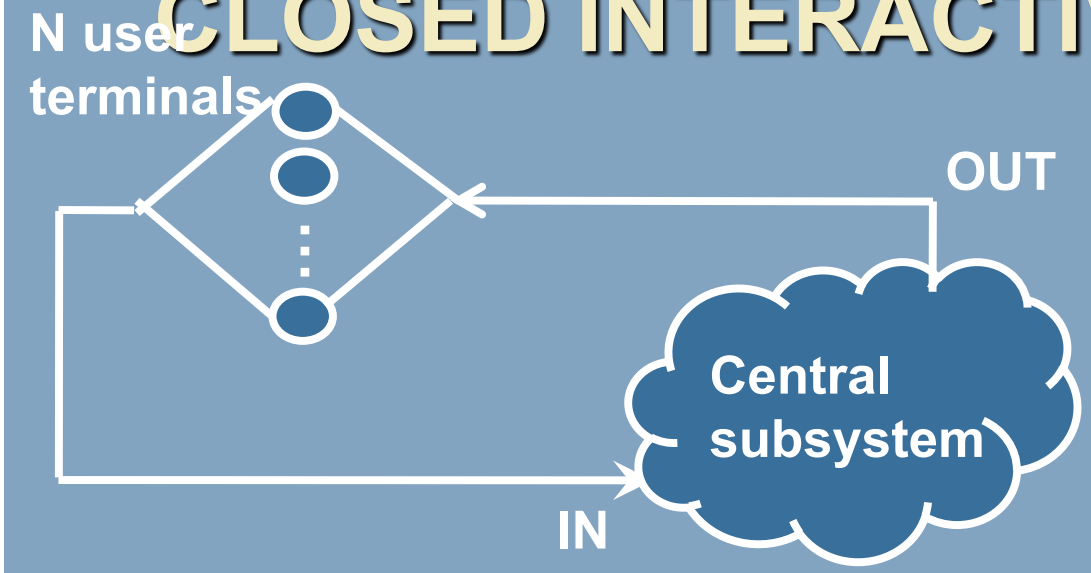
N = number of terminals

Think time Z



- How is response time defined?
  - Time it takes a job to go from IN to OUT

CLOSED INTERACTIVE SYSTEMS



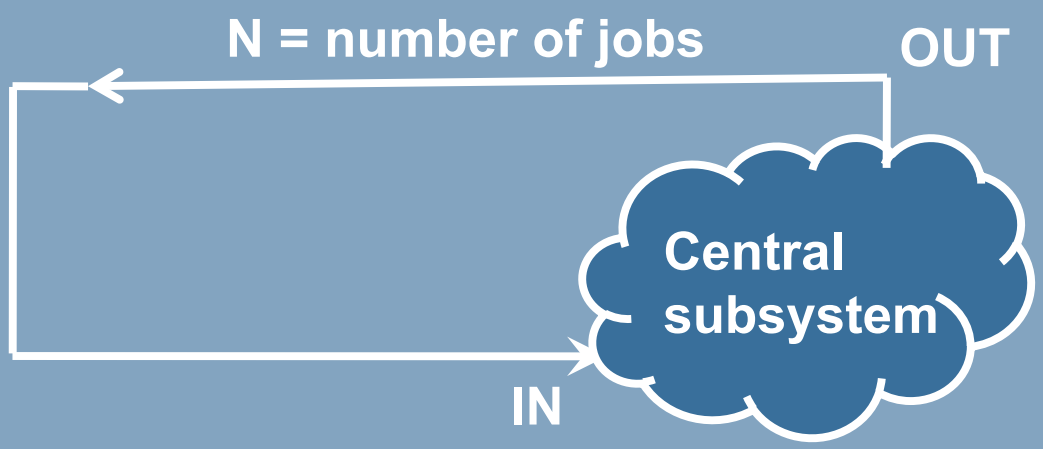
Parameters:

N = number of terminals

Think time Z

- Examples?
- Typical goals:
  - How high can we make N while keeping response times reasonably low?
  - Given a fixed N, what changes to central subsystem will improve response time the most?

CLOSED BATCH SYSTEMS



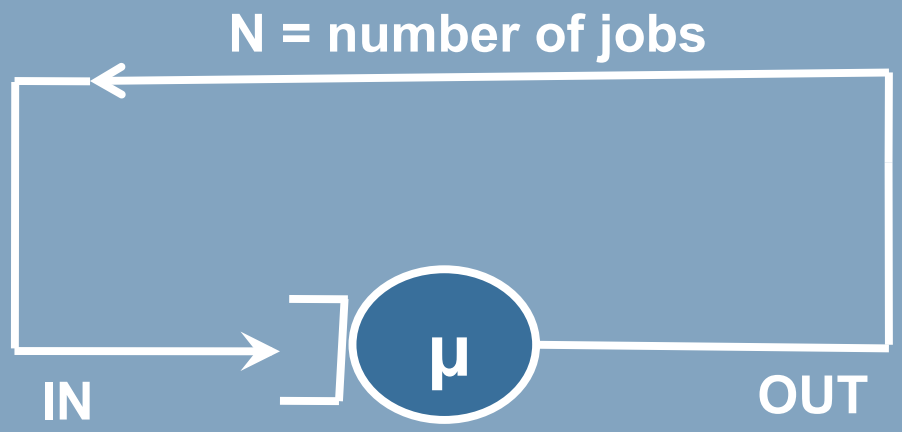
Parameters:

N = number of jobs

- As soon as one job completes, another one is started
- So there's always exactly N jobs in central subsystem

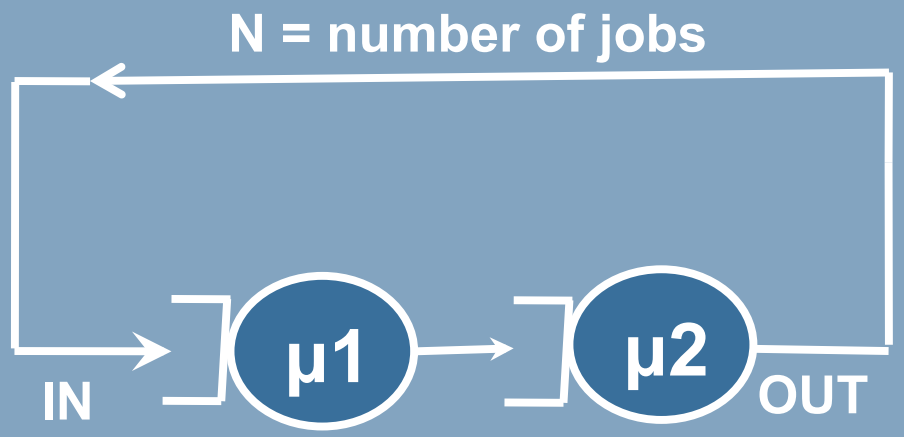
- Typical Goal?
  - Throughput!
- How is throughput defined?

THROUGHPUT IN BATCH SYSTEMS



- What is X?
  - **μ**

THROUGHPUT IN BATCH SYSTEMS



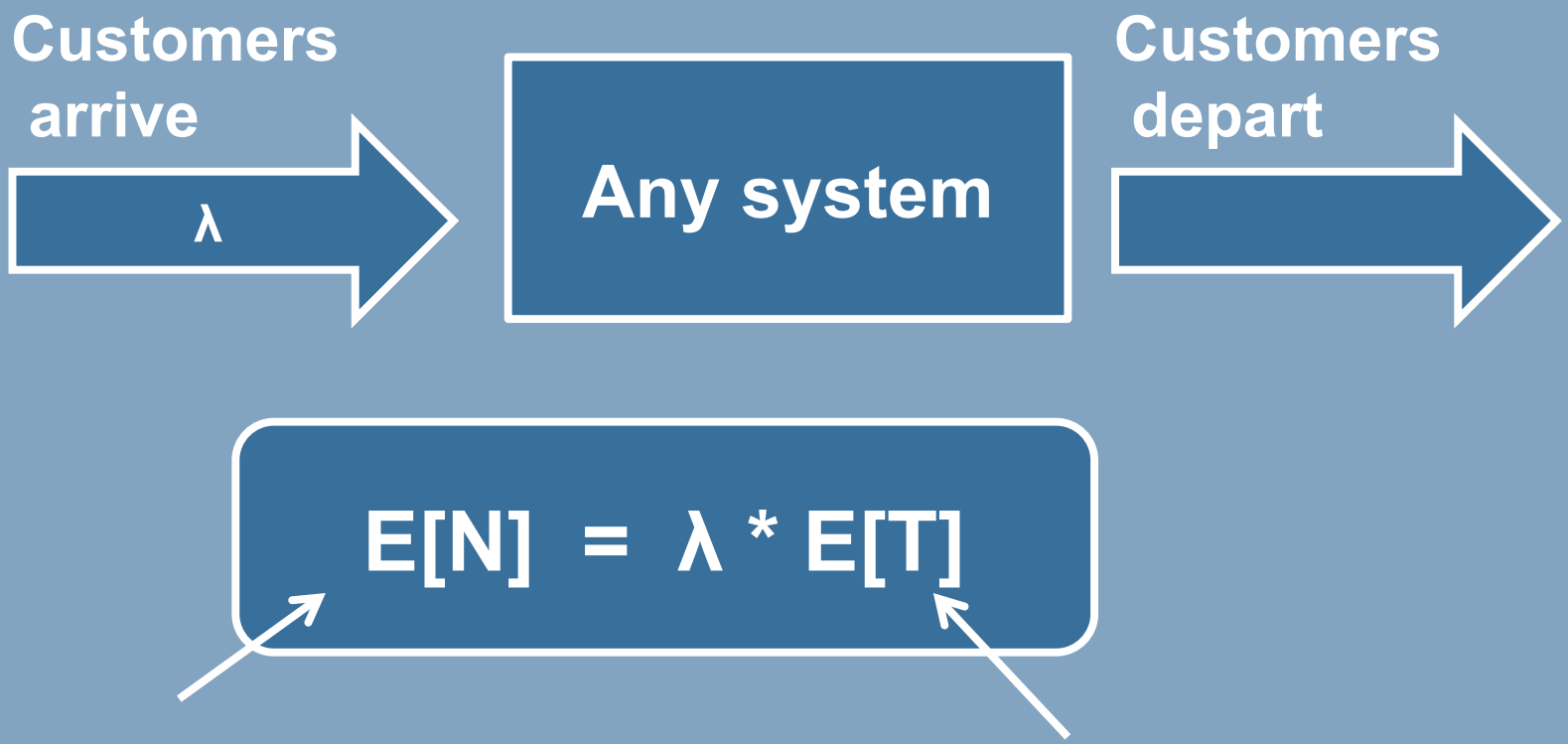
- What is X?
  - It is NOT min **{μ1 ,μ2}**
- **Why?**

SUMMARY SO FAR

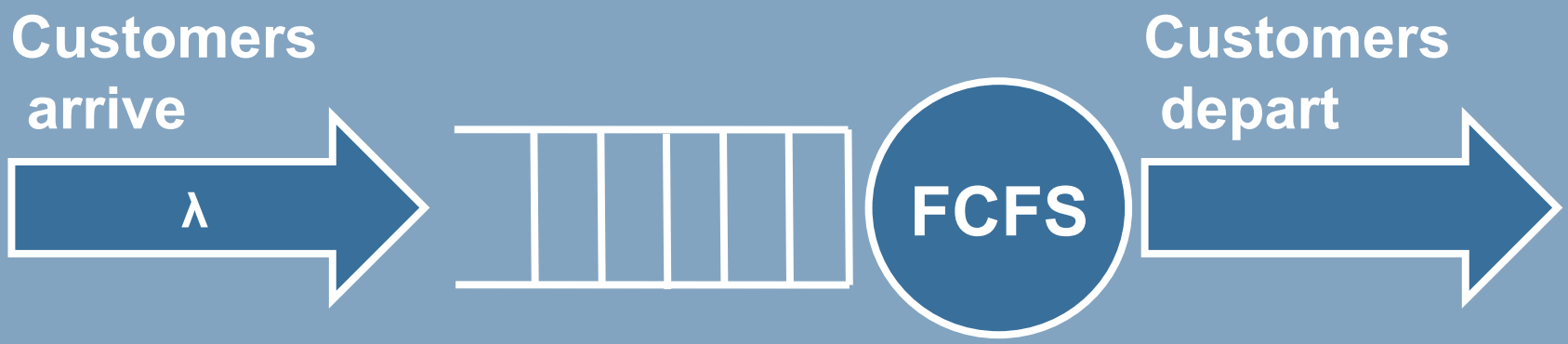
- Open system
  - X independent of **μ**
  - X not affected by doubling **μ**
  - Throughput and response time independent
- Closed system
  - X depends on μ
  - Doubling μ affects X
  - Lower response time => higher X

LITTLE'S LAW

- The single most famous queueing result!
- For open systems:



LITTLE'S LAW: INTUITION



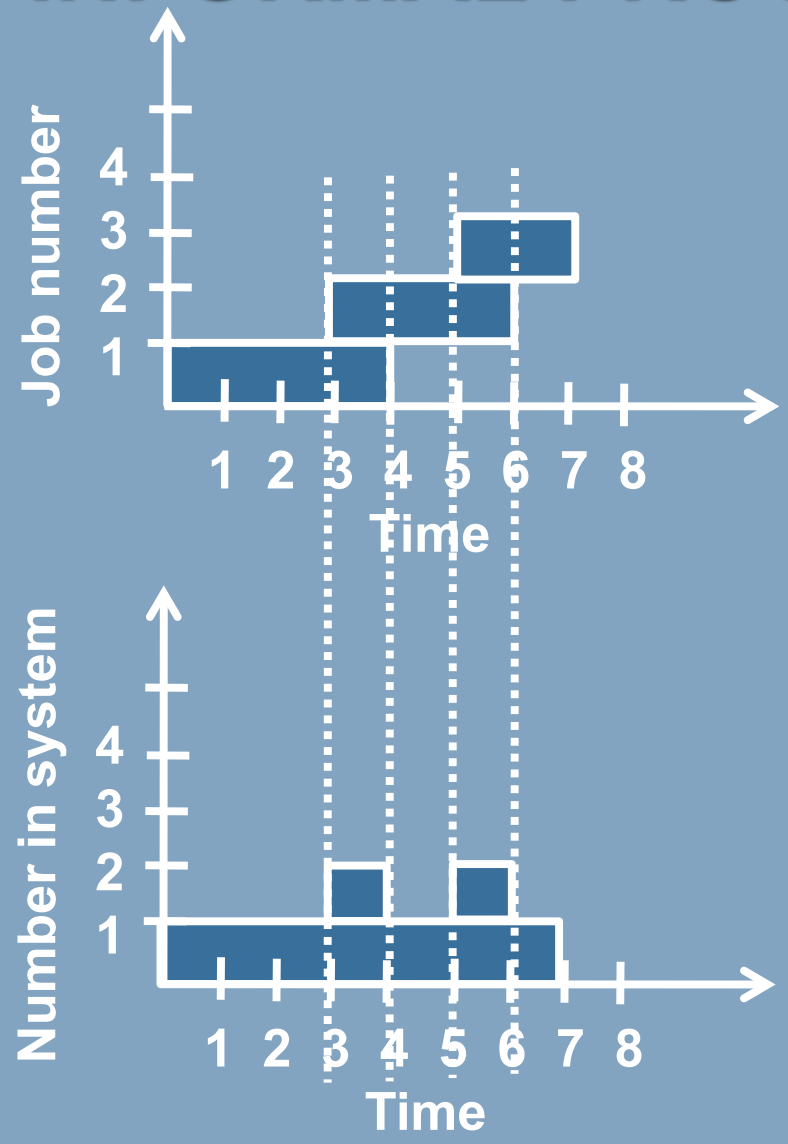
$$E[N] = \lambda * E[T]$$

- Customer arrives, sees E[N] in system
- Average rate of completions is λ
- Expected time until customer leaves:

$$E[T] = E[N] / \lambda$$

LITTLE'S LAW: INFORMAL PROOF

$$E[N] = \lambda * E[T]$$

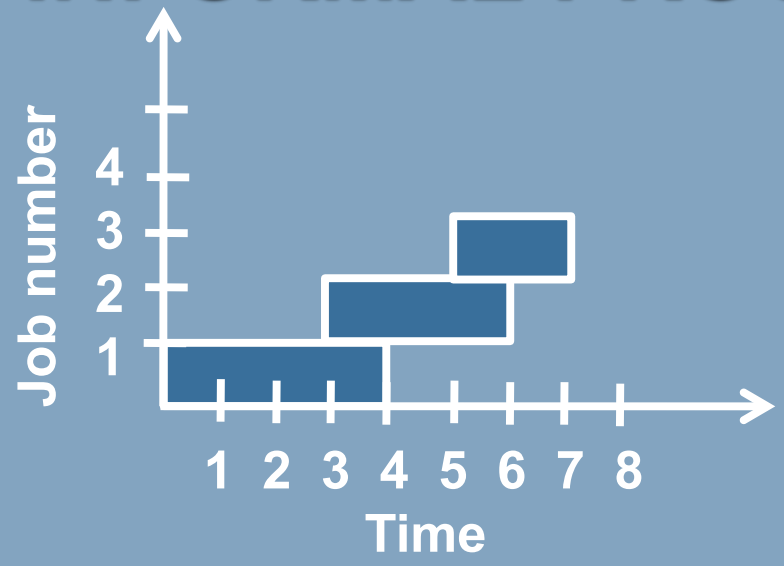


- How does the blue area in graph 1 compare to the blue area in graph 2?

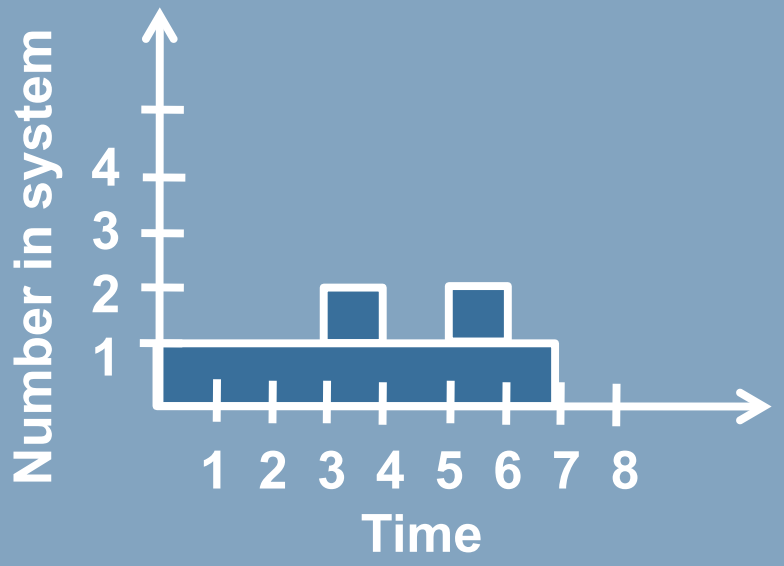
=> they are identical!

LITTLE'S LAW: INFORMAL PROOF

E[N] = λ * E[T]

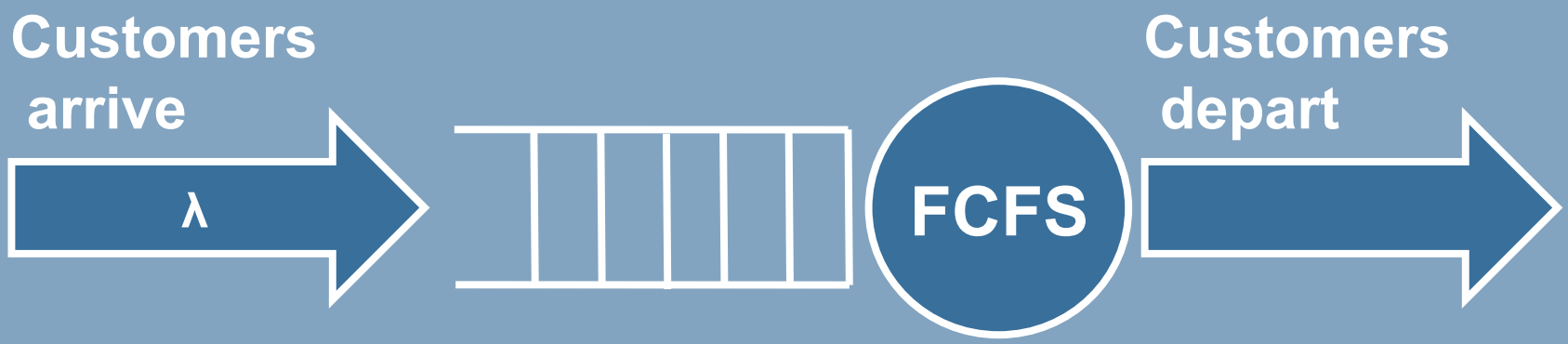


- J = blue area
- n = number jobs completed by time t



- E[T] = J/n
- E[N] = J/t
  = E[T] *n / t
  = E[T] * λ

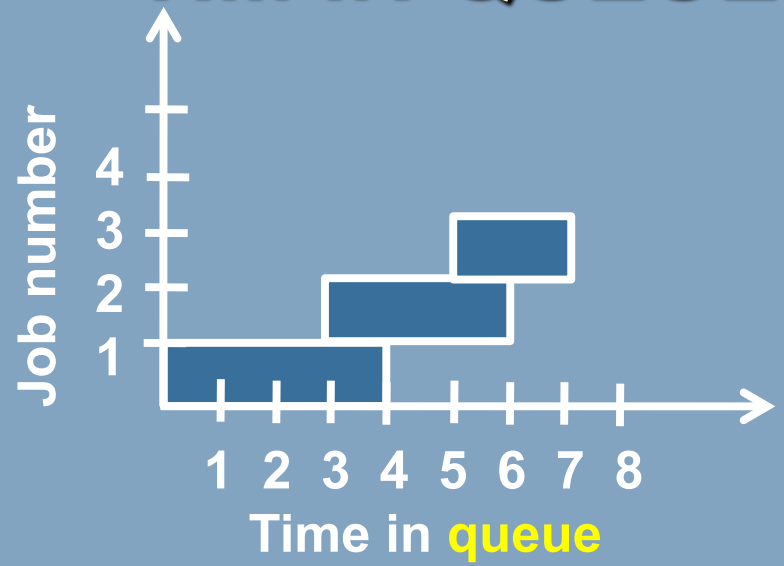
LITTLE'S LAW FOR TIME IN QUEUE

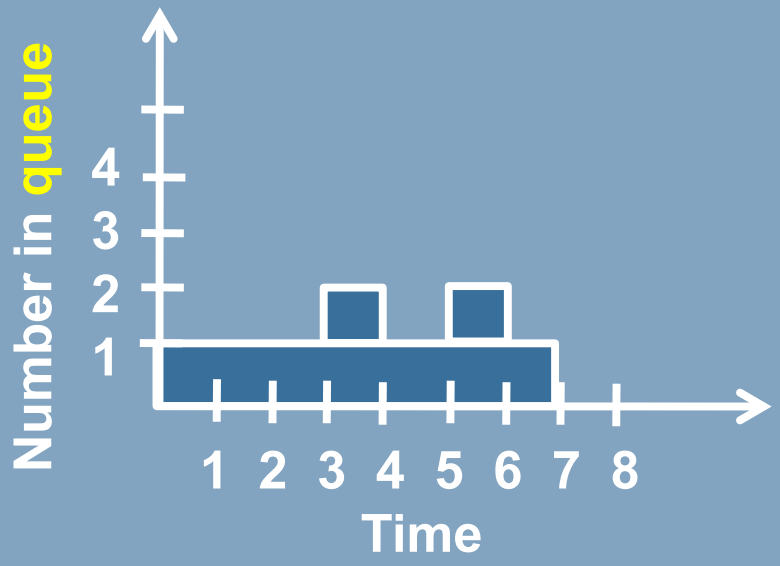


$$E[N_Q] = \lambda * E[T_Q]$$

LITTLE'S LAW FOR TIM IN QUEUE

$$E[N_Q] = \lambda * E[T_Q]$$





- J = blue area
- n = number jobs completed by time t

- E[T_Q] = J/n
- E[N_Q] = J/t
  = E[T_Q] *n / t
  = E[T_Q] * λ

LITTLE'S LAW FOR JOB CLASSES

- Suppose we're interested only in one type of jobs, e.g. "red jobs". Can we apply Little's law to only red jobs?

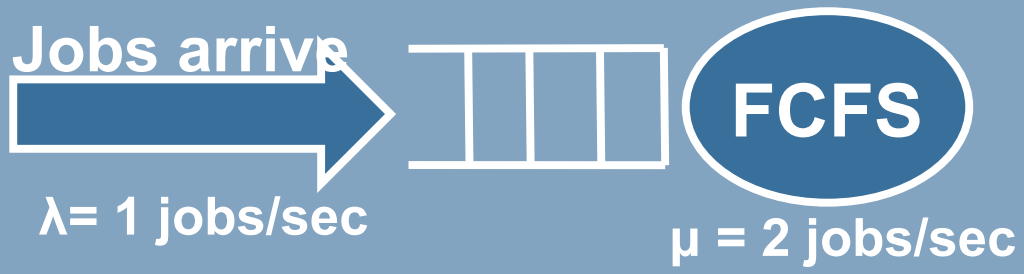
$$E[N_{red}] = \lambda_{red} * E[T_{red}]$$

APPLICATIONS OF LITTLE'S LAW?

Remember our pseudo proof of

$U = \lambda/\mu$?

PSEUDO PROOF OF U = λ/μ



- What is U?
  - Fraction of time server is busy

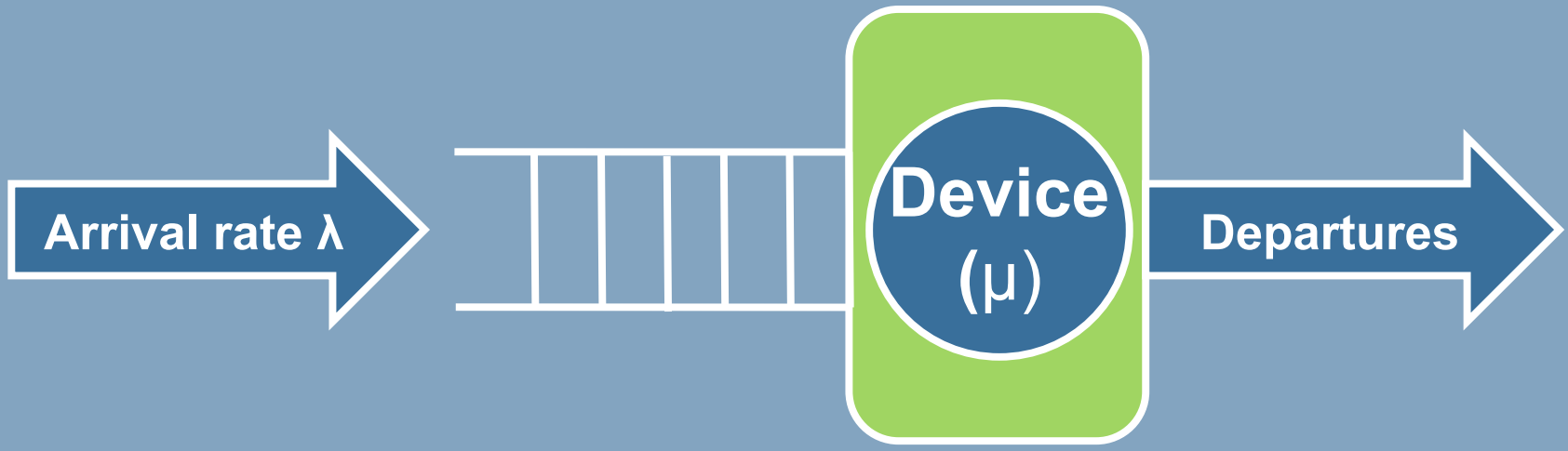
  = mean service time / mean time betw. Arrivals

  = (1/μ)/(1/λ) = λ/μ

APPLICATION OF LITTL

Little's Law:
$E\{N\} = \lambda * E\{T\}$

- Prove that $U = \lambda/\mu$!

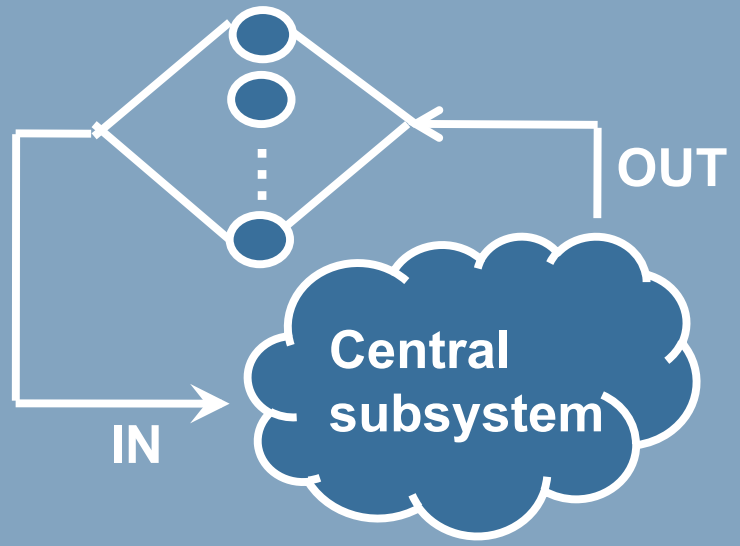


- Consider only the green subsystem
  - What is E{N}? => U
  - What is E{T}? => E[S] = 1/μ
  - **Arrival rate?** => λ

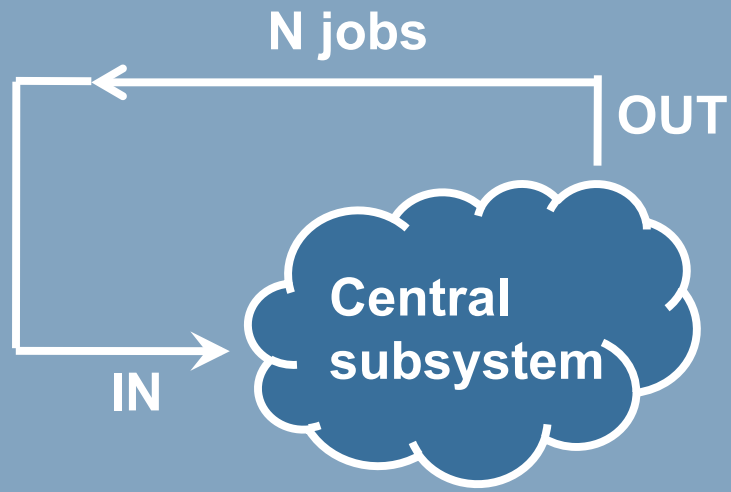
LITTLE'S LAW FOR CLOSED SYSTEMS

- Little's law in open systems: **E{N} = λ * E{T}**
- Little's law in closed systems? **N = X * E{T}**

N users (fixed) with thinktime Z



E[T] = time from "OUT" to "OUT"
E[RespTime] = time from IN to "OUT"



E[T] = time from "IN" to "OUT"

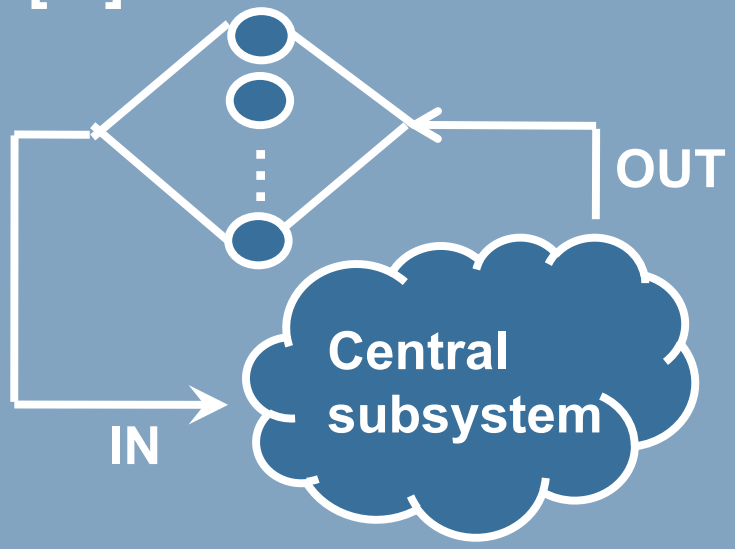
LITTLE'S LAW FOR CLOSED SYSTEMS

- Little's law in open systems: **$E\{N\} = \lambda * E\{T\}$**
- Little's law in closed systems? **$N = X * E\{T\}$**

- Main difference between open & closed?
  - In open, X does not depend on E{T}
  - In closed, X can be improved by improving E{T}

EXAMPLE 1

10 users
E[Z] = 5 sec



E[RespTime] = 15 sec

Little's Law

$$N = X * E\{T\}$$

$$\Rightarrow X = N/E[T]$$

- What is the throughput of this system?
  - What is E[T]? E[T] = E[Z] + E[RespTime]

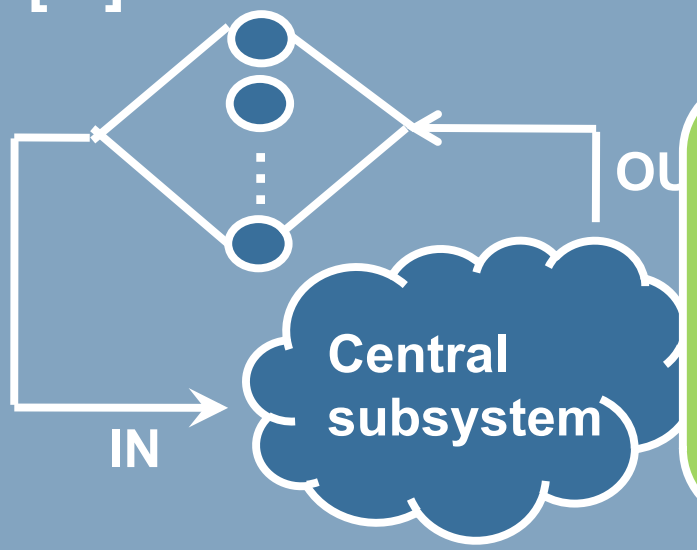
    => X = N / (E[Z]+E[RespTime])

Remember math from last week's paper:

In the flow-based *closed-loop model*, we have a fixed number of users N. Each user goes through a cycle of activity, with a flow of average size S followed by an idle period of average length T_i. The average flow arrival rate is given by $\lambda_c = N/(T_t + T_i)$, where T_t is the average flow completion time. The latter

EXAMPLE 1

10 users
E[Z] = 5 sec



E[RespTime] = 15 sec

Little's Law

Alternative version:

E[R] = N /X -- E[Z]

(Response Time Law)

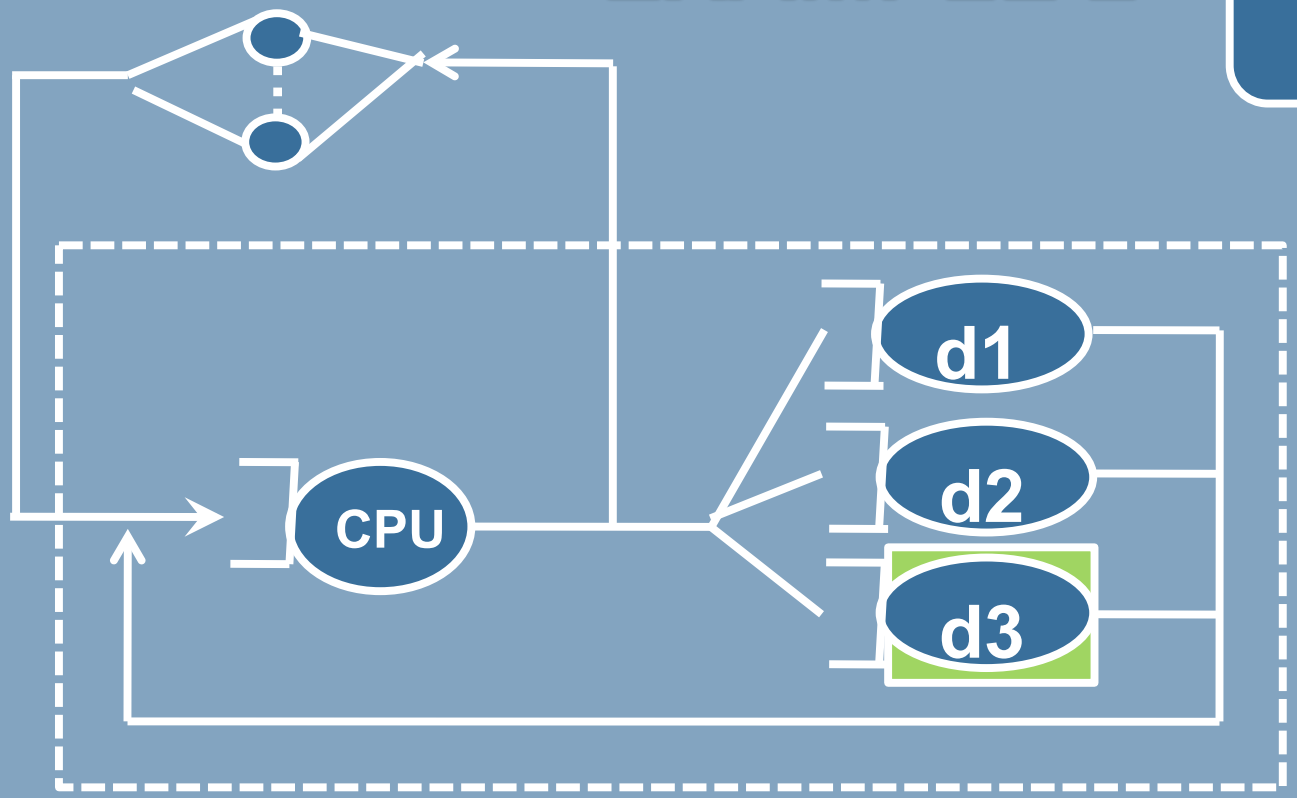
- What is the throughput of this system?
  - What is E[T]? E[T] = E[Z] + E[RespTime]

    => X = N / (E[Z]+E[RespTime])

EXAMPLE 2

Little's Law

$$N = X * E\{T\}$$



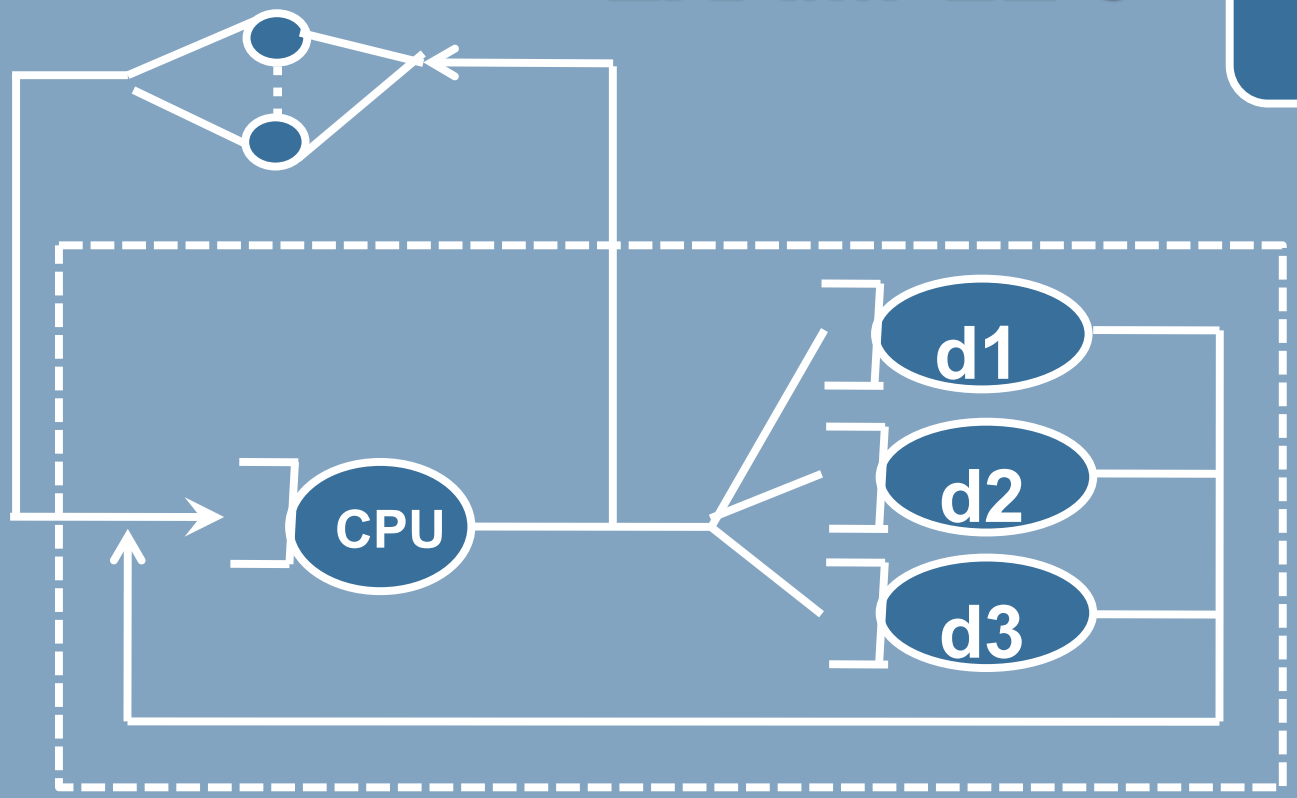
X_d3 = 40 jobs/sec
E[S_d3] = .0225 sec
E[N_d3] = 4

- What is the utilization of disk 3?
- U_d3 = E[number of jobs in green box]
  = X_d3 * E[S_d3] = 90%

EXAMPLE 3

Little's Law

$$N = X * E\{T\}$$



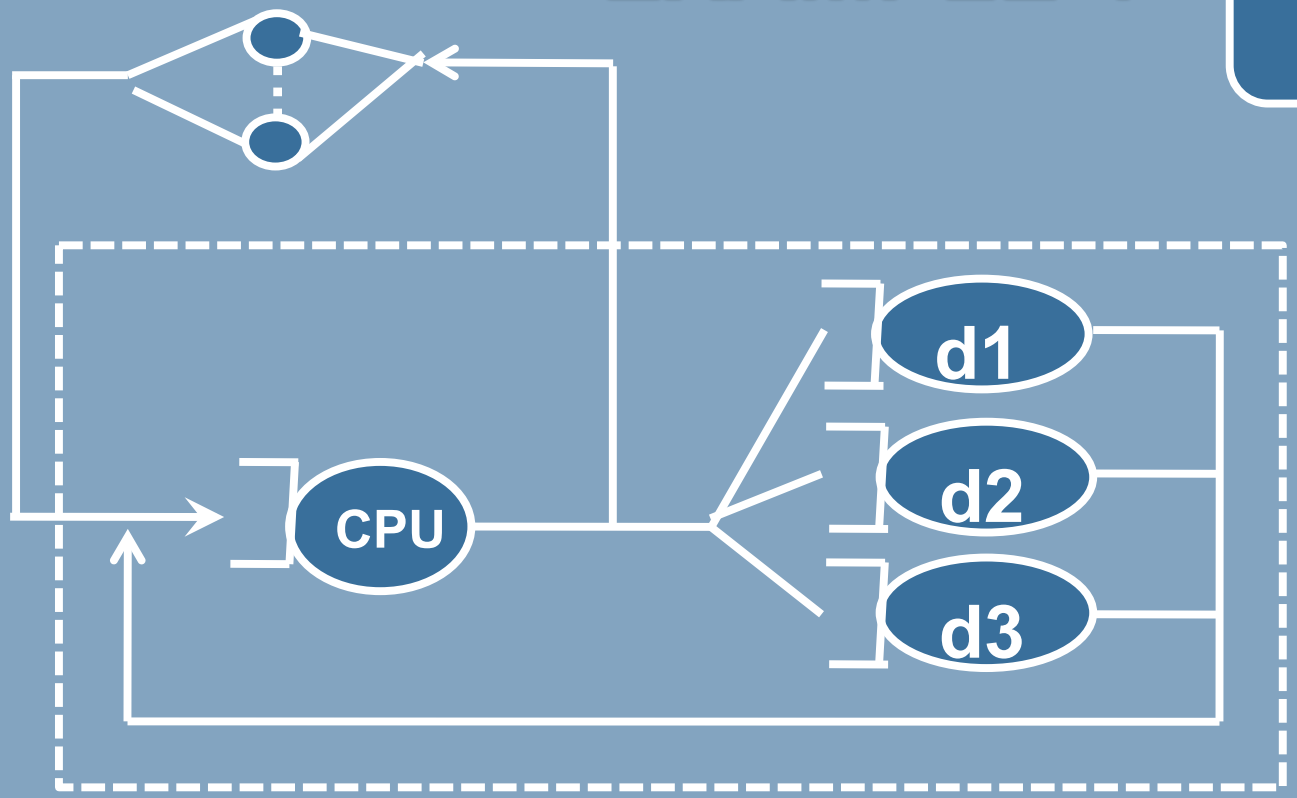
X_d3 = 40 jobs/sec
E[S_d3] = .0225 sec
E[N_d3] = 4

- What is T_Q of disk 3?
- E[T_d3] = N_d3 / X_d3 = 4/40 = 0.1 sec
- E[TQ_d3] = E[T_d3] – E[S_d3] = 0.0775 sec

EXAMPLE 4

Little's Law

$$N = X * E\{T\}$$



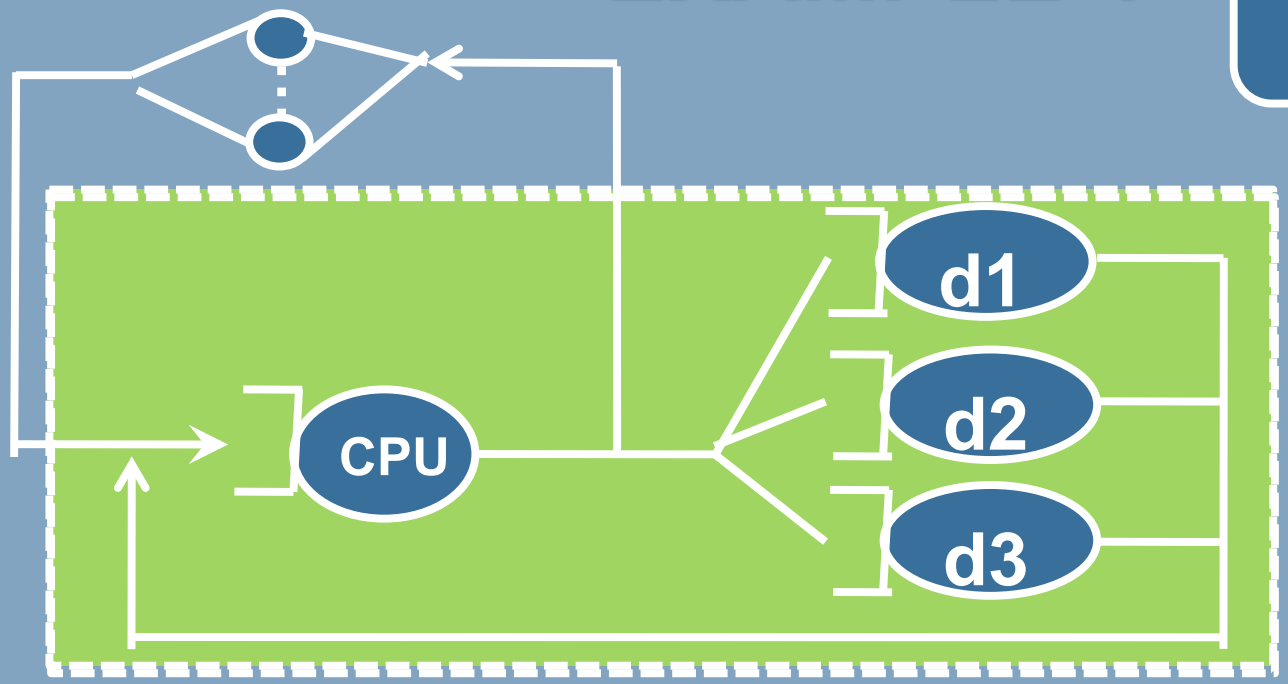
X_d3 = 40 jobs/sec
E[S_d3] = .0225 sec
E[N_d3] = 4

- What is E[number of requests queued at d3] ?
- E[NQ_d3] = E[N_d3] - E[number in service]

  = 4 – 0.9 = 3.1

EXAMPLE 4

Little's Law
$N = X * E\{T\}$



E[ready users] = 7.5
E[Z] = 5 sec

- What is the system throughput ?
  - X = N/ E[T] = N / (E[Z] + E[R]) = 10/(5+E[R])
- But what is E[R]?
  - E[R] = E[N_greenbox] / X_greenbox = 7.5/X

REVIEW

- How to relate the expected number of jobs in the system to expected time spent in the system?

- Open system:

$$E[N] = \lambda * E[T]$$

- Closed batch system (zero thinktime)

$$N = X * E[T]$$

- Closed interactive system

$$E[R] = N/X - E[Z]$$

REVIEW: UTILIZATION LAW

- How can you relate device utilization and device throughput?

$$U_i = X_i * E[S_i]$$

- How can you prove this? Apply Little's law to individual device

$$E[N] = \lambda * E[T]$$