Departamento de Ingeniería Industrial y de Sistemas
Pontificia Universidad Católica de Chile

ICS 3332
Gestión de Operaciones

Clase 20: Programación de Operaciones y Líneas de Espera

Prof. Juan Carlos Ferrer - 2do Semestre 2024

1

Jerarquía en Decisiones de Producción

Pronósticos de Demanda
↓
Planificación Agregada
↓
Programación Maestra de Producción (MPS)
↓
Planificación de Requerimientos de Materiales (MRP)
↓
Programación de Operaciones

2

Problemas de Programación

1. Programación de talleres de trabajo
2. Programación de personal
3. Programación de vehículos
4. Programación de proveedores
5. Programación de proyectos

Programación estática vs. dinámica

3

Programación de talleres de trabajo

- Características del problema
  - Patrón de llegada de los trabajos
  - Número y variedad de máquinas en el taller
  - Número de trabajadores en el taller
  - Evaluación de reglas alternativas

Capacidad

- Objetivos de la programación de talleres
  - Cumplir con las fechas de entrega
  - Minimizar el tiempo promedio de producción
  - Minimizar el inventario en proceso (WIP)
  - Alta utilización de máquinas y trabajadores
  - Reducir los tiempos de cambio
  - Minimizar costos de producción y trabajadores

Servicio al cliente

Eficiencia de la planta

4

Reglas de secuenciamiento

- FIFO (*First In First Out*) o FCFS (*First Come First Serve*)
  - Trabajos son procesados en el orden en que llegan al taller
- SPT (*Shortest Processing Time*)
  - Trabajos son ordenados en orden creciente por sus tiempos de proceso. Luego, el trabajo con el menor tiempo de proceso es el primero.
- EDD (*Earliest Due Date*)
  - Trabajos son ordenados en orden creciente de sus fechas de entrega. El trabajo con la fecha más cercana se hace primero.
- CR (*Critical Ratio*)
  - Balance entre SPT y EDD, ya que considera el cuociente entre tiempo restante para fecha de entrega y tiempo de proceso. El trabajo con menor tasa crítica es primero.

5

Ejemplo: Comparación de reglas

6

Líneas (colas) de espera

"La otra fila siempre avanza más rápido"

"Si te cambias de fila, la que dejaste comenzará a avanzar más rápido que la actual"

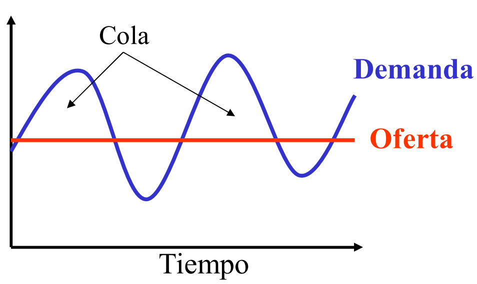
Las personas destinan 2-5 años de sus vidas esperando en colas.

(Fuente: Time.com 2007)

7

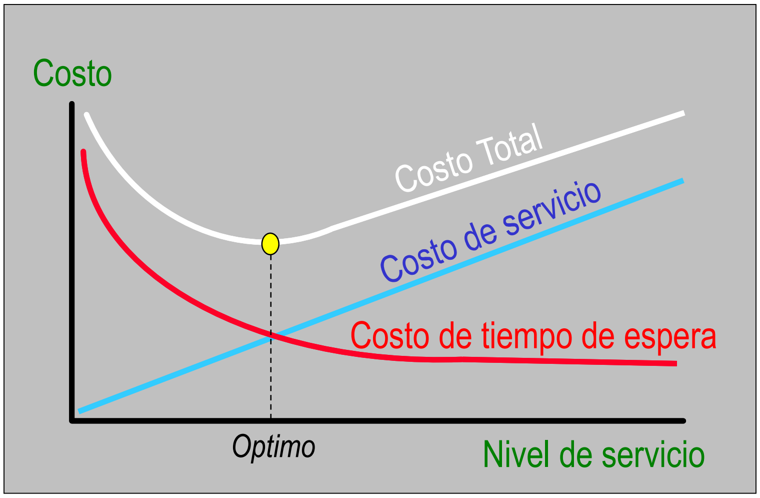
Líneas de espera

- A. K. Erlang in 1913 fue el primero en estudiarlas
  - Analizó centrales telefónicas
- Área de investigación llamada "Teoría de colas"
- Problema de decisión
  - Balancear el costo de proveer un buen servicio, con el costo del tiempo de espera de los clientes



8

Costos en líneas de espera



9

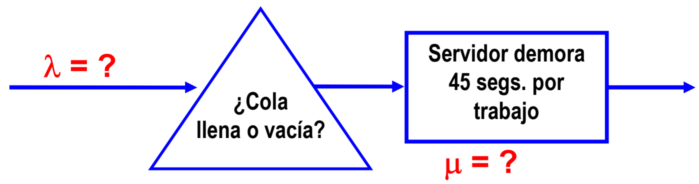
Ejemplos

Situación	Llegadas	Servidor	Servicio
Banco	Clientes	Cajero	Depósito, etc.
Hospital	Pacientes	Doctor	Tratamiento
Intersección de tráfico	Autos	Semásforo	Tráfico controlado
Línea de ensamble	Partes	Trabajador	Ensamble

11

Veamos las siguientes situaciones

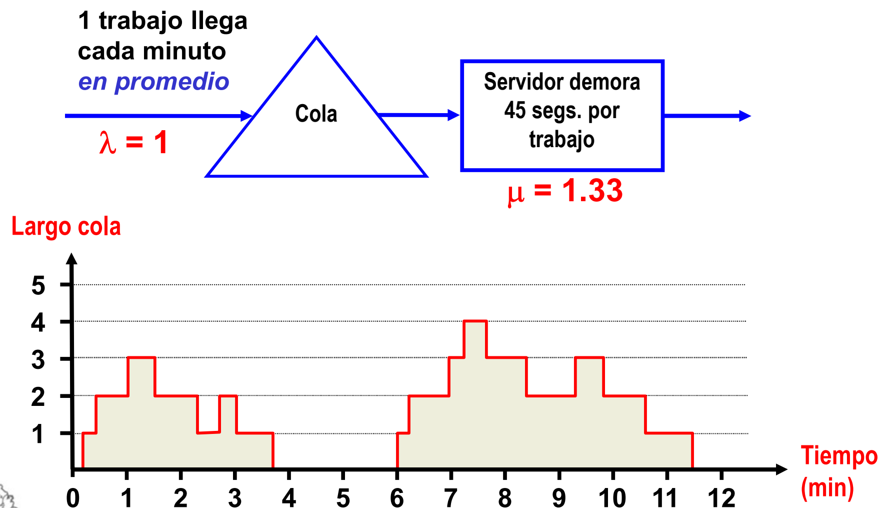
- Llega un trabajo cada un minuto



- El próximo trabajo llega:
  - Después de 15 segs con probabilidad 0.5
  - Después de 1 min 45 segs con probabilidad 0.5

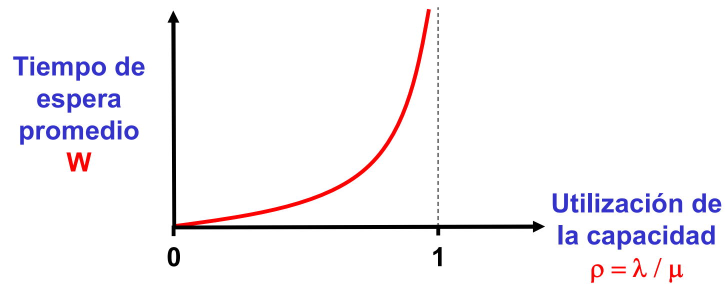
12

Incertidumbre genera las colas



13

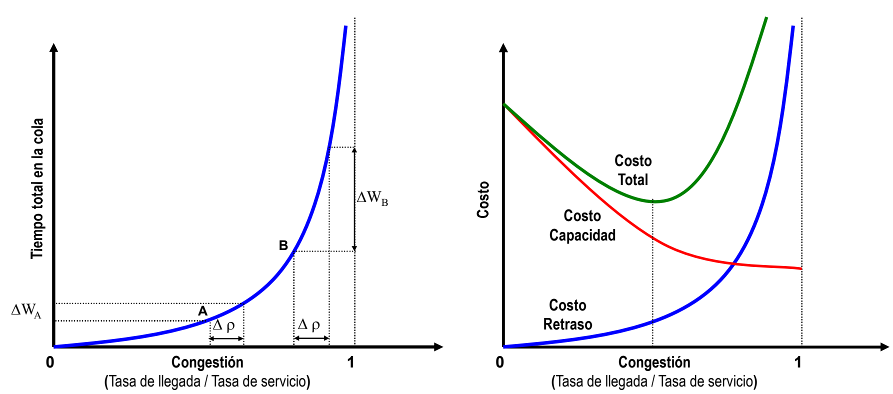
Intuición principal en líneas de espera



La relación entre tiempo de espera y utilización de la capacidad es fuertemente no lineal

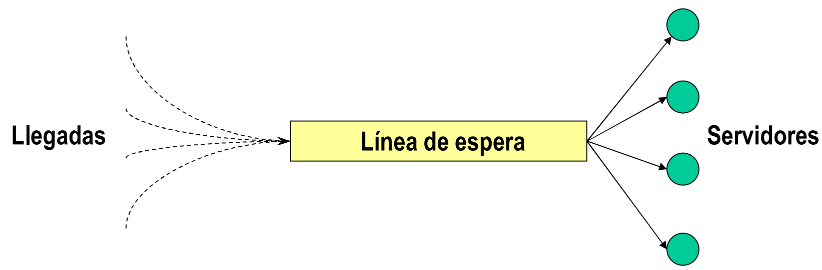
14

Manejando la física de las colas



15

Características de líneas de espera



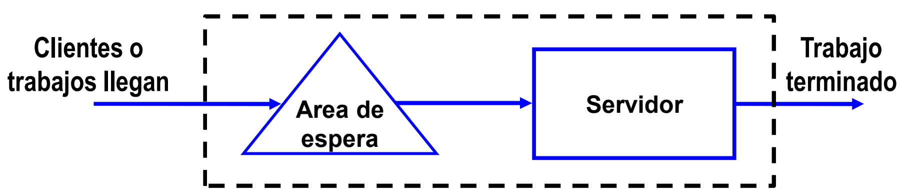
- Población de clientes puede ser finita o infinita
- Patrón de llegada puede ser random
- Clientes pueden ser o no ser pacientes

- Largo de la cola puede ser limitado
- Clientes son servidos en una secuencia FIFO
- Pueden haber más de una cola

- Tiempos de servicio pueden ser random
- Puede haber más de un servidor

16

Medidas de desempeño



Desempeño del sistema = F(Parámetros del sistema)

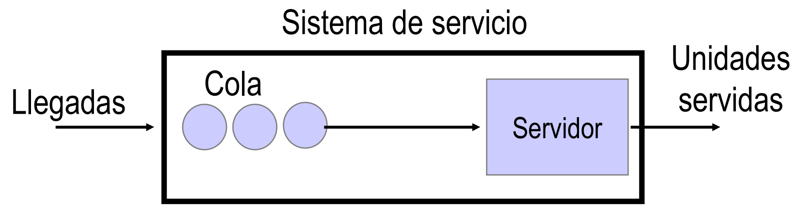
γ Output/Throughput rate
L Nivel inventario / largo cola
W Tiempo de espera
C Tiempo de ciclo
P_{full} Probabilidad de cola llena

λ Tasa de llegada
μ Tasa de servicio
M Tiempo de servicio
S Número de servidores
R Capacidad de la cola (buffer)
ρ Utilización de capacidad
K Clases de servicio
π Política de servicio

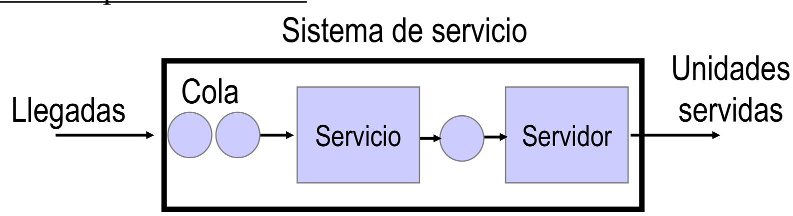
17

Tipos de líneas de espera

Canal Simple – Una fase



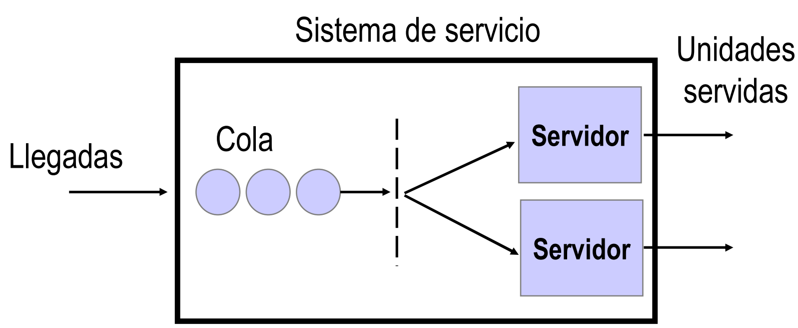
Canal Simple – Multi-fase



19

Tipos de líneas de espera

Multi-canal – Una fase

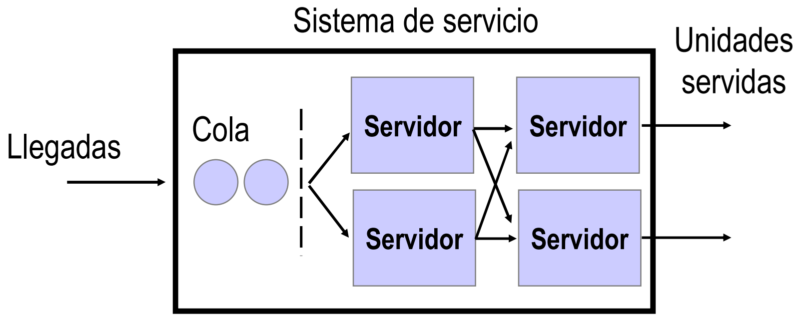


Ejemplo: banco, correo, etc.

20

Tipos de líneas de espera

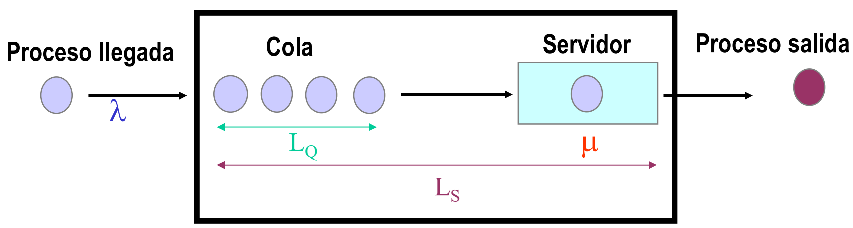
Multi-canal – Multi-fase



Ejemplo: Lavandería, sistema de producción make-to-order, etc

21

Analicemos una cola simple (1 servidor)

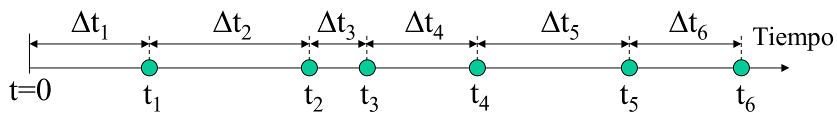


Cantidades importantes:

- Tasa llegada: λ (ej: clientes/hr, paquetes/seg, etc)
- Tasa servicio: μ (ej: clientes/hr, etc)
- # unidades en el sistema: L_S
- # unidades en la cola: L_Q
- Tiempo en el sistema: W_S (Throughput Time)
- Tiempo en la cola: W_Q
- Utilización del servidor: ρ

22

Cola simple: Proceso de llegada



Si la tasa de llegada es λ (clientes/hr) entonces **en promedio** cada hora llegan λ clientes al sistema.

Ejemplo: Supongamos que llegan 6 clientes en promedio por hora (λ=6 client/hr). Por lo tanto se espera que llegue un cliente cada 10 minutos.

¿Qué son los (Δt)?

Si en promedio llegan λ clientes en una hora, entonces **en promedio** un cliente llega cada 1/λ horas

23

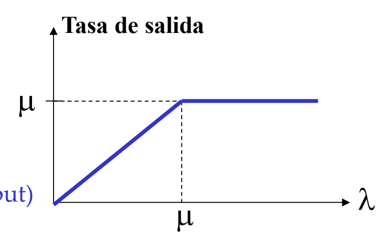
Cola simple: Proceso de servicio

Si la tasa de servicio es μ (clientes/hora) entonces **en promedio** el servidor es capaz de procesar μ clientes cada hora.

Ejemplo: Supongamos que la tasa de servicio es μ=10 (clientes/hr). En promedio, ¿cuántos clientes completarán el servicio cada hora si:

- λ=0?
- λ=5?
- λ=10?
- λ=15?
- λ=20?

Tasa salida = min(λ , μ)
Output efectivo (throughput)



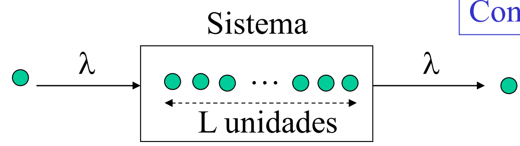
Si la tasa de servicio es μ clientes por hora, entonces **en promedio** el servidor necesita 1/μ horas para servir cada cliente.

24

Cola simple: Principio de conservación

Ley de Little

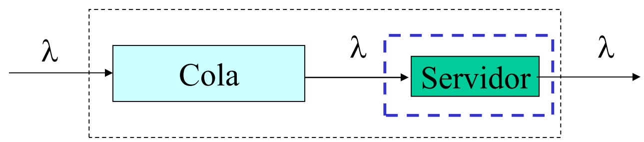
Condición: $\mu > \lambda$



Throughput Time (Tiempo de espera) = W

$$\boxed{L = \lambda W}$$

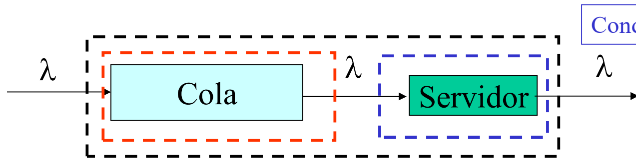
Aplicación: Utilización del servidor (ρ)



25

Cola simple: Relaciones importantes

Condición: $\mu > \lambda$



$L_S = \lambda W_S$

$L_Q = \lambda W_Q$

$\rho = \lambda / \mu$

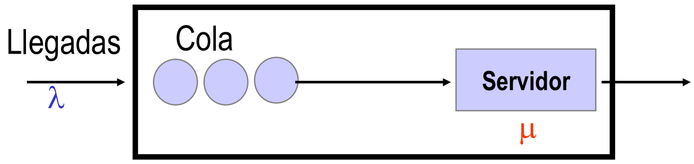
$L_S - L_Q = \lambda (W_S - W_Q) = \lambda$ * (tiempo servicio) $= \lambda \times 1/\mu = \rho$

$W_S = W_Q + 1/\mu$

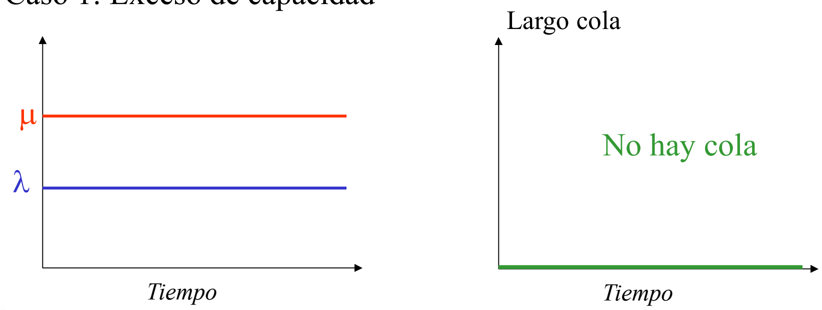
Fracción del tiempo que el servidor está ocioso = 1- ρ

26

Modelo determinístico - FIFO



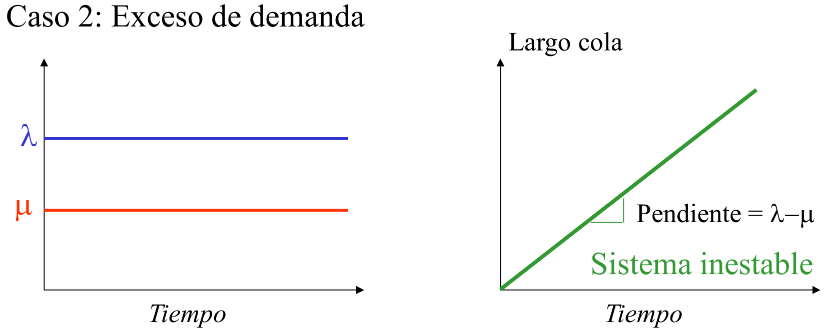
Caso 1: Exceso de capacidad



27

Modelo determinístico - FIFO

Caso 2: Exceso de demanda

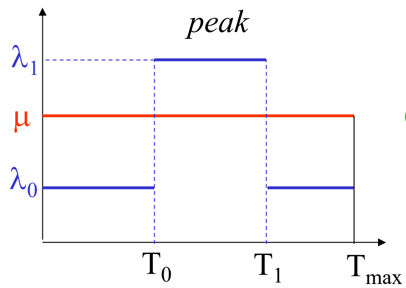


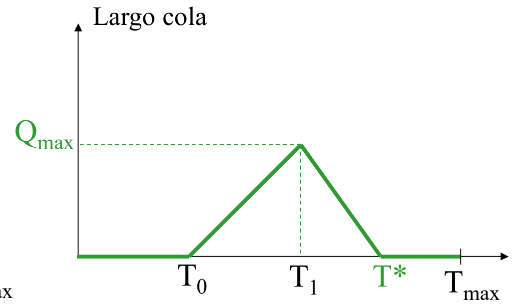
<u>Observación:</u> Estabilidad en un sistema de colas requiere que la tasa de producción (capacidad) sea más grande que la tasa de la demanda, es decir, $\lambda < \mu$.

28

Modelo determinístico - FIFO

Caso 3: Demanda *peak*





Algunas preguntas

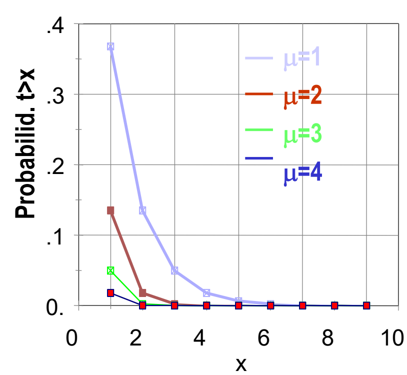
- ¿Cual es el valor de Q_{max} y T*?
- ¿Cual es el largo promedio de la cola durante el periodo $[0,T_{max}]$?
- ¿Cual es el tiempo promedio de espera por orden en $[0,T_{max}]$?

29

Cola simple: M/M/1

- Un servidor
- Llegadas Poisson
- Tiempos de servicios exponenciales

$$\Pr(Tiempo > x) = e^{-\mu x}$$



30

Cola simple: M/M/1 - Ecuaciones

 # promedio de unidades en el sistema $L_s = \frac{\lambda}{\mu - \lambda}$

Tiempo promedio en el sistema $W_s = \frac{1}{\mu - \lambda}$

 # promedio de unidades en la cola $L_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$

Tiempo promedio en la cola $W_q = \frac{\lambda}{\mu(\mu - \lambda)}$

Utilización del sistema $\rho = \frac{\lambda}{\mu}$

31

Psicología de líneas de espera

1. Tiempo ocioso se percibe más largo que tiempo ocupado
2. Esperas inciertas parecen más largas que las conocidas
3. Esperas inexplicadas parecen ser más largas que aquellas con explicación
4. Esperas injustas se hacen más largas que esperas justas
5. Mientras más valioso el servicio, más tiempo esperará el cliente
6. La ansiedad hace que las esperas parezcan más largas
7. Esperas individuales se perciben más largas que esperas grupales

32

Leer

- "*Perspectives on Queues: Social Justice and the Psychology of Queueing*" by R. Larson
- "*Queueing Theory*" by V.G. Narayanan

33