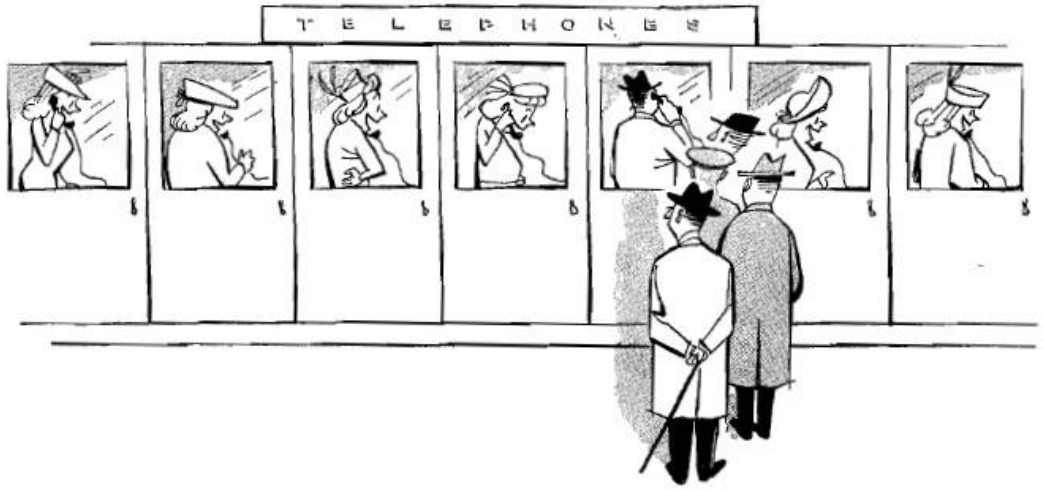
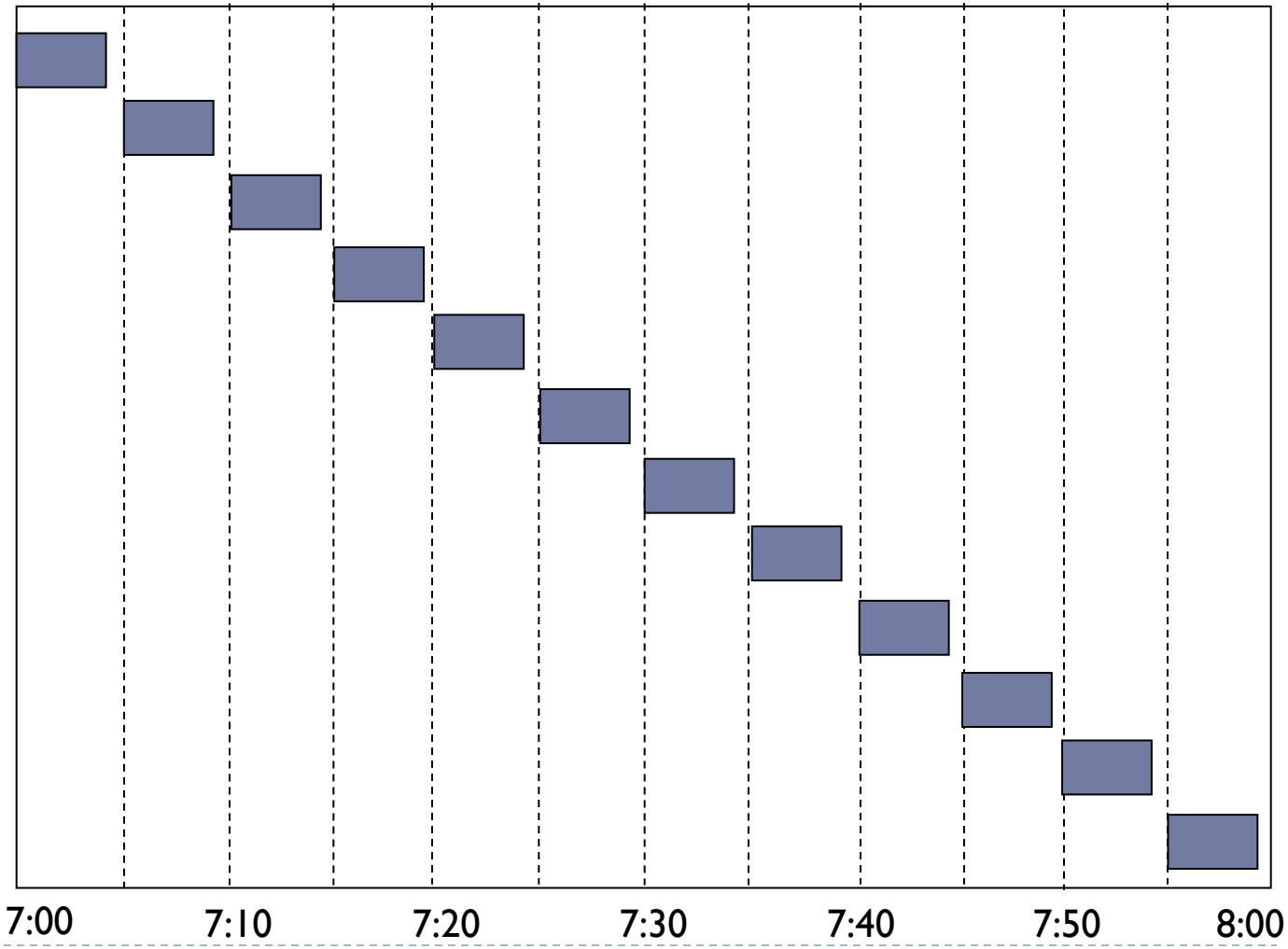


Administrado la Variabilidad y los Tiempos de Espera

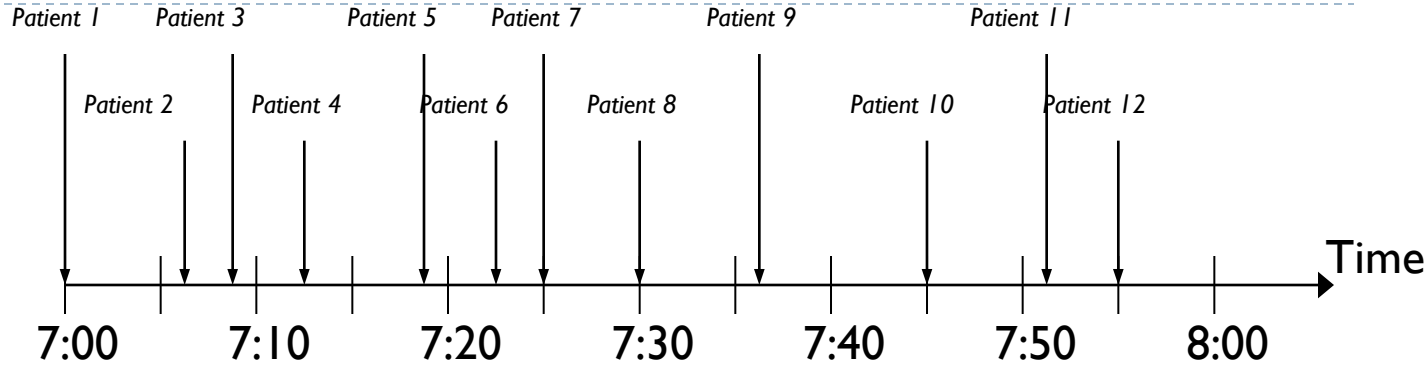
Un Proceso Extraño

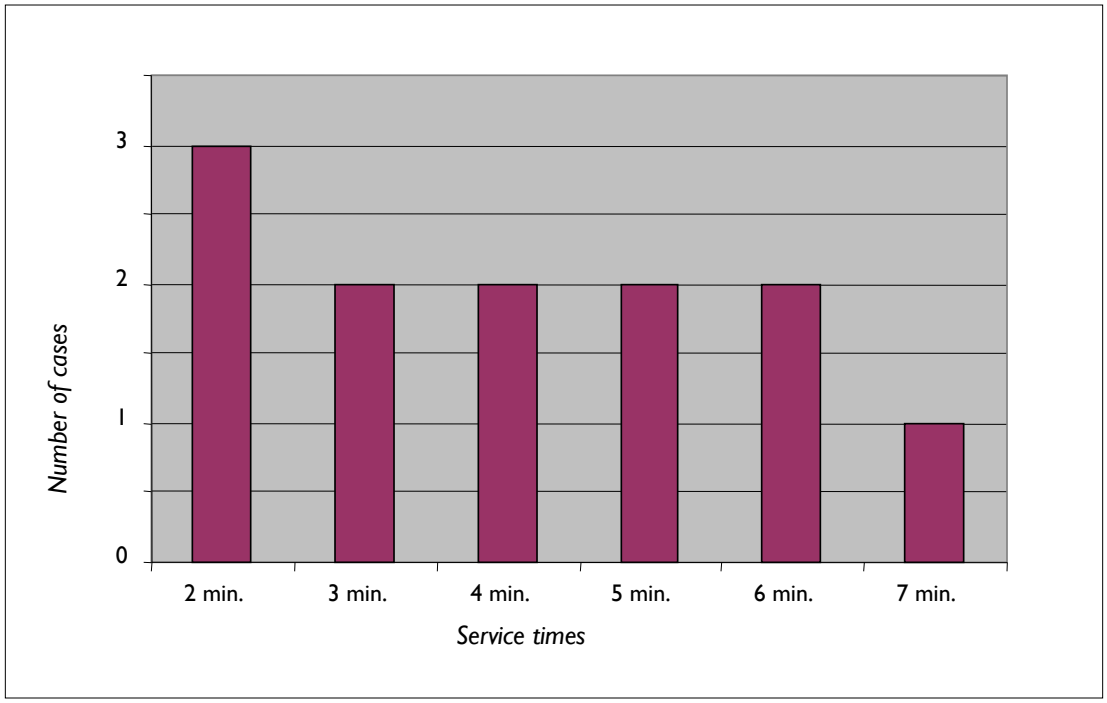
Patient	Arrival Time	Service Time
1	0	4
2	5	4
3	10	4
4	15	4
5	20	4
6	25	4
7	30	4
8	35	4
9	40	4
10	45	4
11	50	4
12	55	4



Un Proceso más Realista

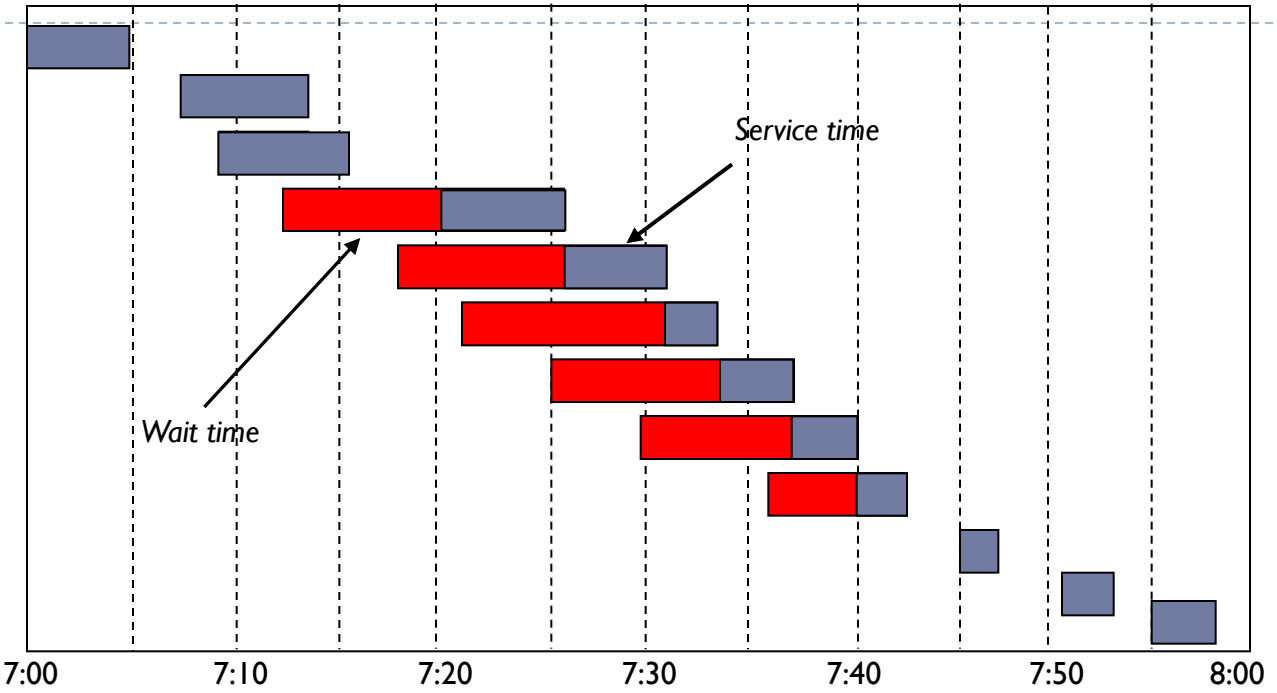
Patient	Arrival Time	Service Time
1	0	5
2	7	6
3	9	7
4	12	6
5	18	5
6	22	2
7	25	4
8	30	3
9	36	4
10	45	2
11	51	2
12	55	3



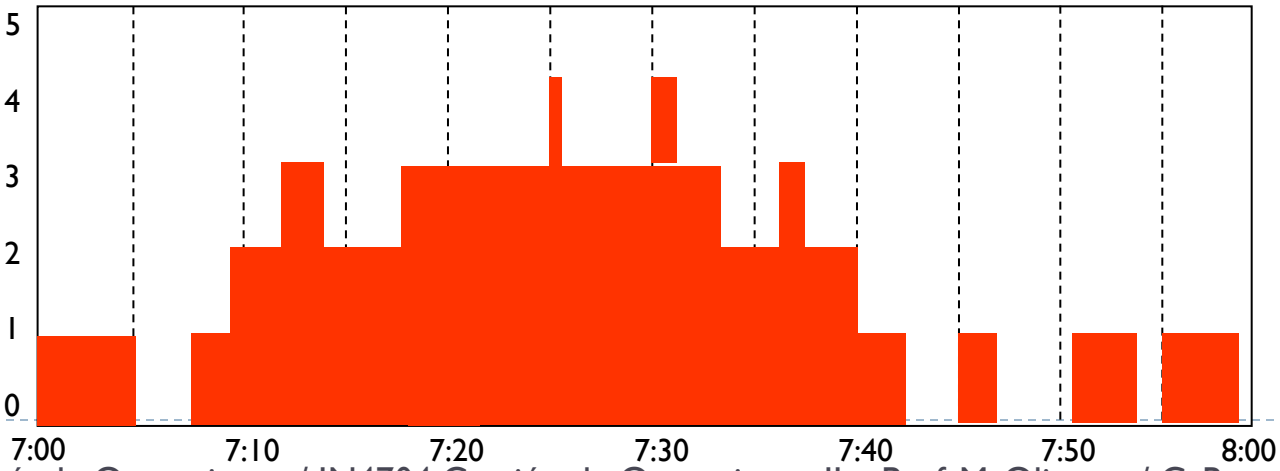


La variabilidad lleva a tiempos de espera e Inventario

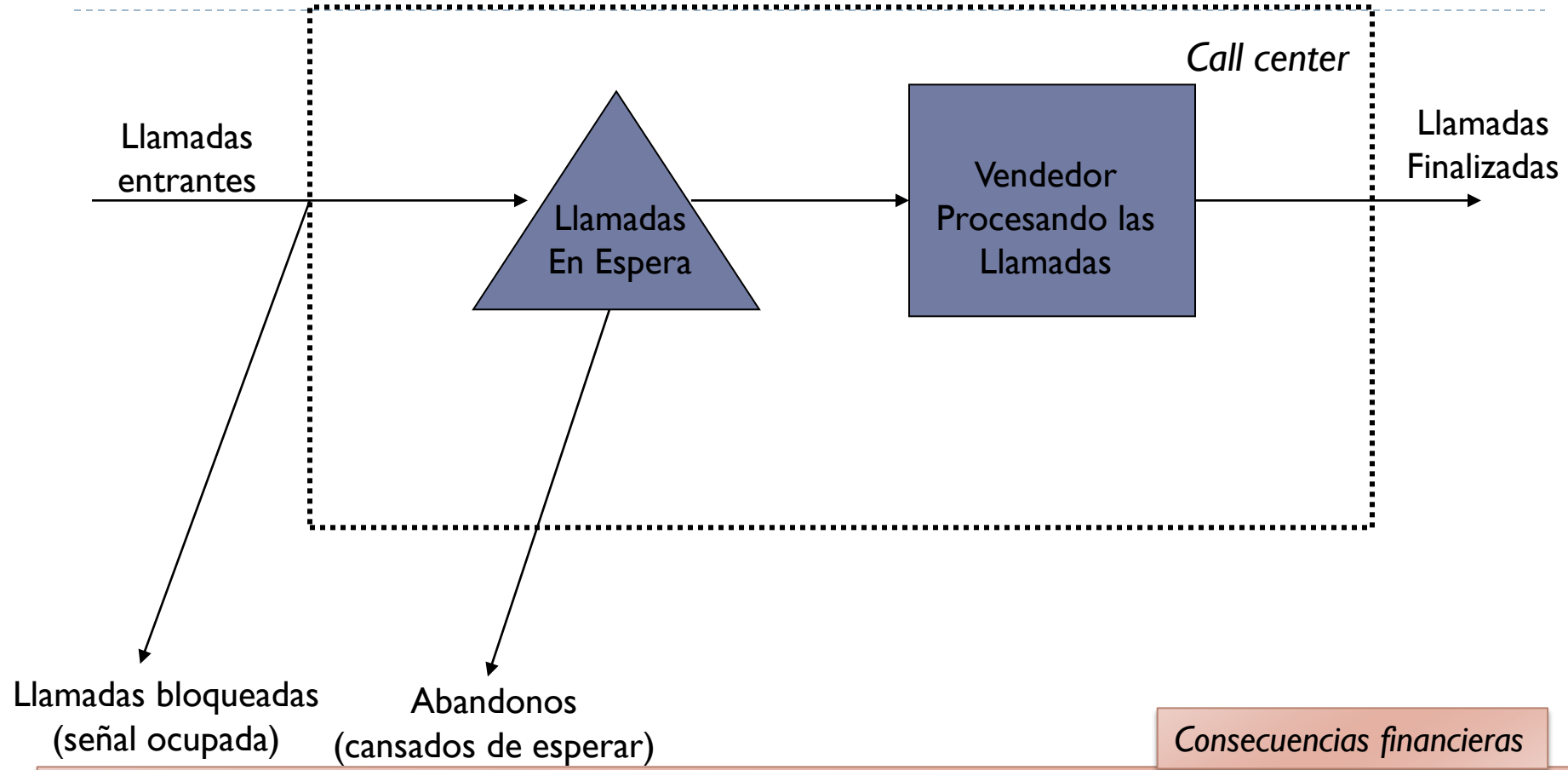
Patient	Arrival Time	Service Time
1	0	5
2	7	6
3	9	7
4	12	6
5	18	5
6	22	2
7	25	4
8	30	3
9	36	4
10	45	2
11	51	2
12	55	3



Inventory (Patients at lab)



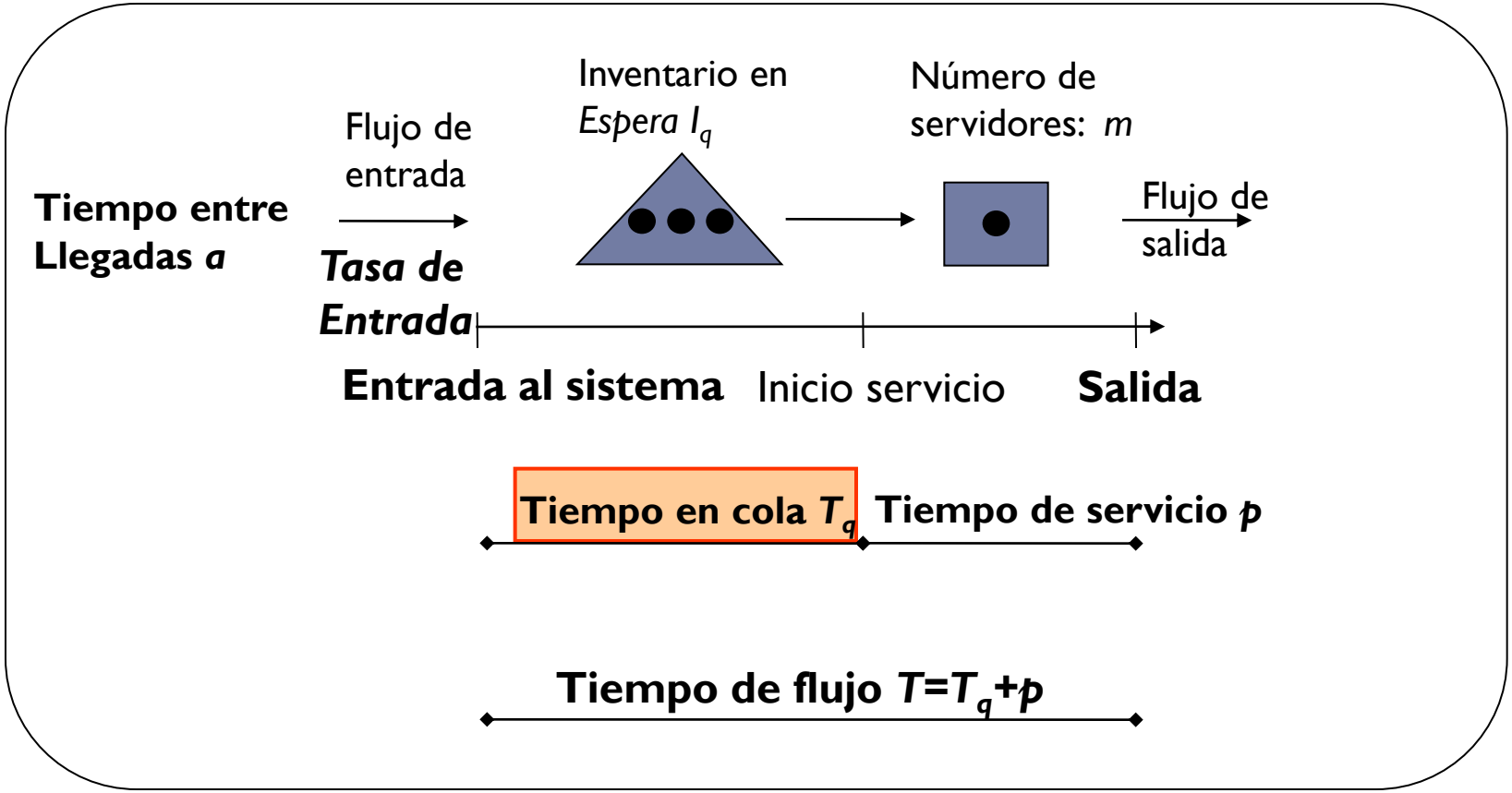
Ejemplo de un Sistema de Colas: Call Center



Consecuencias financieras

Pérdida de flujo	Costos de espera Pérdida de imagen Flujo perdido (abandonado)	Costo de capacidad Costo por cliente	$$$ Ganancia $$$

Modelo de Espera en un Sistema



Métricas del Nivel de Servicio

- Muchas organizaciones definen un estándar de servicio como un *Tiempo de Espera Aceptable*.
  - Es un límite superior para el tiempo experimentado por un porcentaje dado (grande) de clientes, o *Nivel de Servicio*.

- AWT= *Tiempo de Espera Aceptable*= Máximo tiempo de espera en la cola experimentado por SL% de los clientes.

- SL = *Nivel de Servicio* = Porcentaje de clientes cuyo tiempo de espera es menor o igual al AWT.

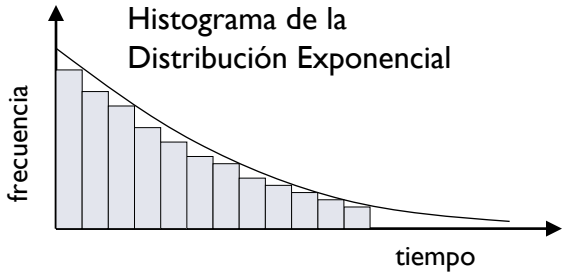
Métricas del Nivel de Servicio (II)

- Ejemplo: muchos call centers son diseñados de tal manera que tengan un SL = 80% o 90% para un AWT de 20 segundos.
  - La gran parte de los acuerdos con call centers sub-contratados especifican un *Acuerdo de Nivel de Servicio (SLA, Service Level Agreement)* de este tipo.
- Otras métricas:
  - T_q = Tiempo de espera promedio en la cola.
  - T_s = Tiempo de espera promedio en el sistema. (WTS)
  - N_q = Número promedio de clientes en la cola = (tasa de demanda)* Tq (Fórmula de Little)
  - N_s = Número promedio de clientes en el sistema = (tasa de demanda) * Ts (Fórmula de Little)
  - P_d= Probabilidad de demora = Posibilidad de que un cliente deba esperar.

Modelando la Llegada y el Tiempo de Servicio

- Para incorporar la variabilidad, un modelo de colas generalmente requiere un descripción detallada de la distribución estadística de llegadas y tiempos de servicio.

  Ejemplo:
  Tiempo entre llegadas consecutivas tiene una distribución exponencial (llegadas de Poisson)



- *Coeficiente de variación (CV):* mide la variabilidad de la variable aleatoria X.

$$CV_X = \left(\frac{\text{Desv. Est(X)}}{\text{Prom (X)}} \right)$$

CV_a: tiempos de llegada

CV_p: tiempos de servicio

Ejemplo:
Si el tiempo entre llegadas consecutivas sigue una distribución exponencial, entonces CVa = 1 (propiedad particular de esta distribución).

El Modelo Erlang o M/M/S

Supuestos básicos:

(a) Existe un grupo de agentes de servicio (servidores) con habilidades y características idénticas.

(b) Los clientes se sirven de forma FIFO, no existen prioridades.

(c) El servicio y los tiempos entre llegadas son aleatorios. El modelo Erlang exacto asumo que ambos tienen una distribución específica (exponencial).

  (a) La distribución exponencial está caracterizada completamente por un único parámetro, su media (y distribución estándar).

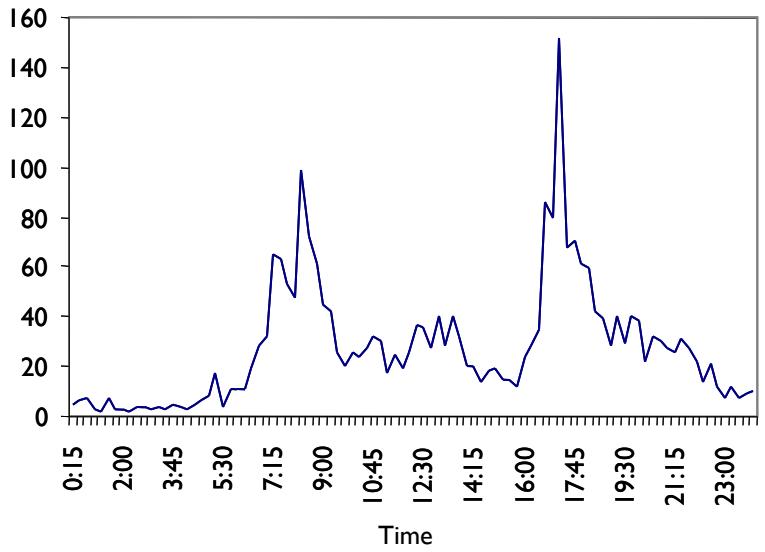
  (b) En la práctica, el modelo Erlang se usa como una aproximación, incluso si este supuesto no se cumple.

(d) El espacio de espera es amplio (hay capacidad suficiente).

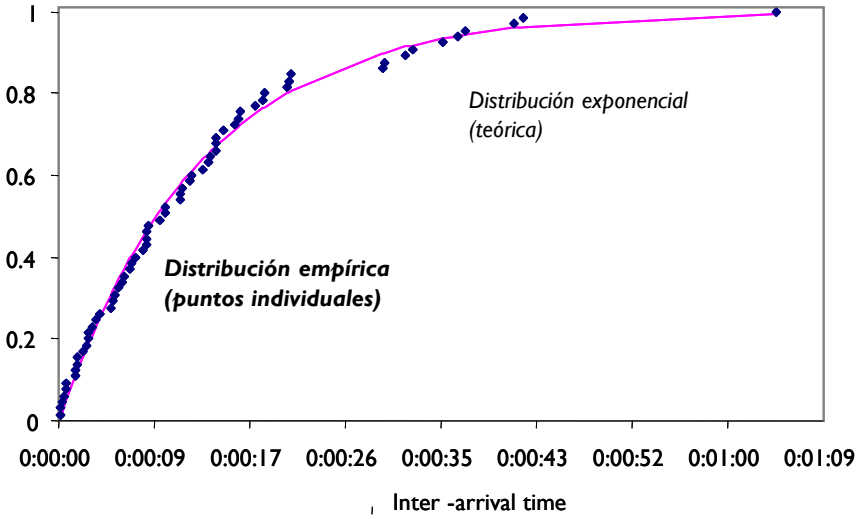
(e) Los clientes no se van del sistema antes de ser servidos, no hay abandonos.

Datos en un Call Center real

Cantidad de clientes cada 15 minutos.



Función de Distribución



- Estacionalidad v/s variabilidad
- Se necesita dividir en intervalos los datos

- Dentro de cada intervalo, el tiempo entre llegadas consecutivas sigue una distribución exponencial.

Carga ofrecida: Tasa de Utilización

- Sea

s = Número de agentes / servidores.

a = λ/μ = carga ofrecida = número mínimo de agentes requeridos.

ρ = a/s= tasa de utilización

- Bajo cualquier tipo de aleatoriedad, se debe tener que $s > a$ o $\rho < 1$.
- La diferencia $(s - a)$ puede ser pensada como la *capacidad basada en el servicio*. Su magnitud dependerá del nivel de servicio que se desea o se necesita proveer.

Modelo Erlang: Fórmulas Básicas

Nivel de Servicio:

$$SL = 1 - P_d(s,a)\, e^{-(s-a)\,AWT\mu} = 1 - P_d(s,\rho)\, e^{-s(1-\rho)\,AWT\mu} \qquad (1)$$

Tiempo de espera promedio:

$$T_q = \frac{P_d(s,a)}{\mu(s-a)} = \frac{P_d(s,\rho)}{\mu s(1-\rho)} \qquad (2)$$

Probabilidad de demora:

$$P_d(s,a) = \frac{a^s / s!}{[1-\rho]\left[\sum_{k=0}^{s-1} a^k/k! + \frac{a^s}{s^!}\frac{1}{(1-\rho)}\right]} \qquad (3)$$

Modelo Erlang: Probabilidad de demora

Propiedades de la probabilidad de demora P_d:

a) P_d aumenta de 0 a 1 a medida que ρ aumenta de 0 a 1.
b) Para s=1, $P_d(1,a) = \rho = a$
c) Para $s \geq 10$: $P_d(s,a) \approx \Phi(\frac{s-a}{\sqrt{a}})$, con $\Phi(\cdot)$ la distribución normal estándar.
d) Para valores dados de a y s, el nivel de servicio proveído no depende en el tiempo <u>absoluto</u> de espera estándar (AWT), sino que el tiempo de espera <u>relativo:</u>

$$\mathrm{AWT} * \mu = \mathrm{AWT} / \left(\frac{1}{\mu}\right)$$

= Tiempo de espera estándar, expresado como porcentaje del tiempo de servicio promedio.

Medidas de servicio cuando s=1

Probabilidad de demora

P (no demora) = $1-\rho$

==> P(demora) = 1 - P(no demora) = ρ

Tiempo promedio de espera en la cola

$$T_q = \frac{1}{\mu}\frac{\rho}{1-\rho}$$

Tiempo promedio pasado en el <u>sistema</u>

$$T_s = T_q + \frac{1}{\mu} = \frac{1}{\mu-\lambda}$$

Medidas de servicio cuando s=1 (II)

Largo promedio de la cola

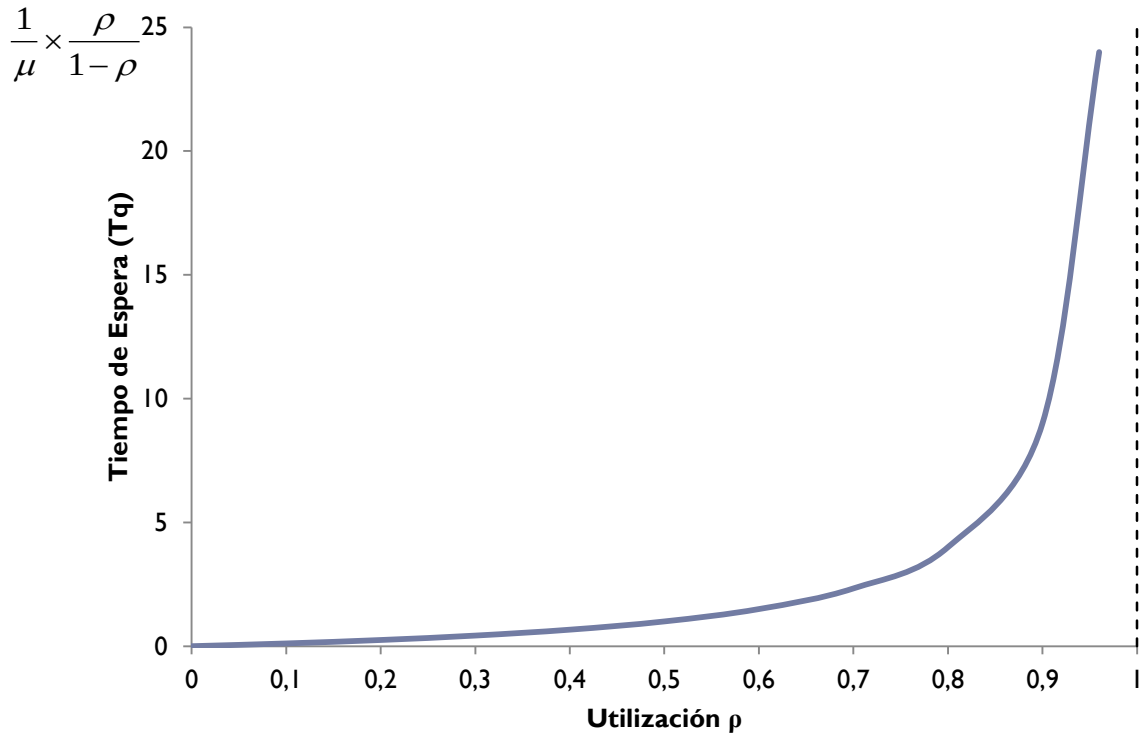
Little's Law =>

$$N_q = \lambda T_q = \frac{\rho^2}{1-\rho}$$

Número promedio de clientes en el sistema

$$N_s = \lambda T_s = \frac{\lambda}{\mu\lambda} = \frac{\rho}{1-\rho}$$

Utilización y Espera Promedio

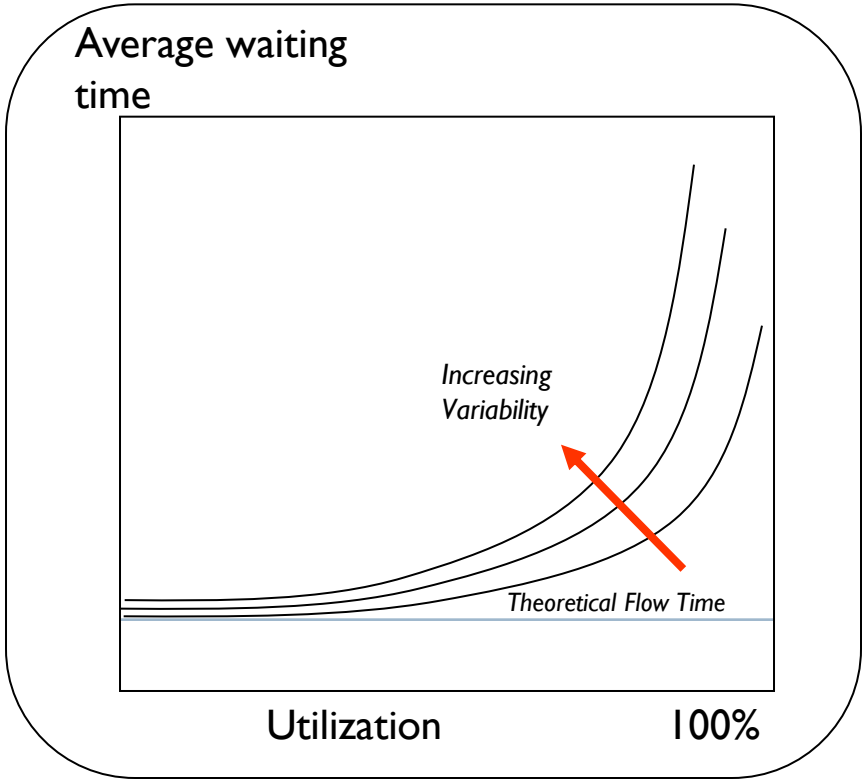


Servicio y Tiempos de Llegada No-Exponenciales

- La variabilidad en el servicio y los tiempos entre llegadas son quienes impulsan el tiempo de espera.
- Midiéndolo con el *Coeficiente de Variación:*

$$CV_s = \frac{\text{Desv. Est. del tiempo de servicio}}{\text{Tiempo de servicio promedio}}$$

$$CV_a = \frac{\text{Desv. Est. del t' entre llegadas}}{\text{Tiempo prom. entre llegadas}}$$



- Aproximación del tiempo de espera promedio en la cola:

$$\underbrace{T_q,\ \text{Erlang}}_{\text{Tiempo de espera en cola Erlang}} \cdot \underbrace{\left(\frac{CV_a^{\,2} + CV_s^{\,2}}{2}\right)}_{\text{Factor de variabilidad en el t' de servicio}}$$

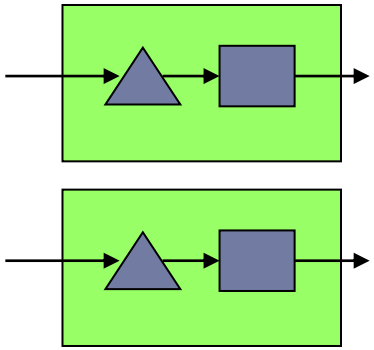
Ejemplo: Retailer Online

Los clientes envían preguntas a un retailer online a través de un chat de mesa de ayuda online cada dos minutos en promedio, y la desviación estándar de los tiempos entre llegadas también es de dos minutos. El retailer online tiene tres empleados para responder preguntas. Toma un promedio de cuatro minutos escribir una respuesta. La desviación estándar del tiempo de servicio es de dos minutos.

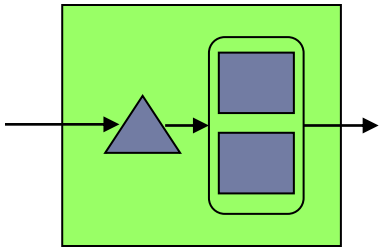
P: Estimar la espera promedio del cliente antes de ser servido.

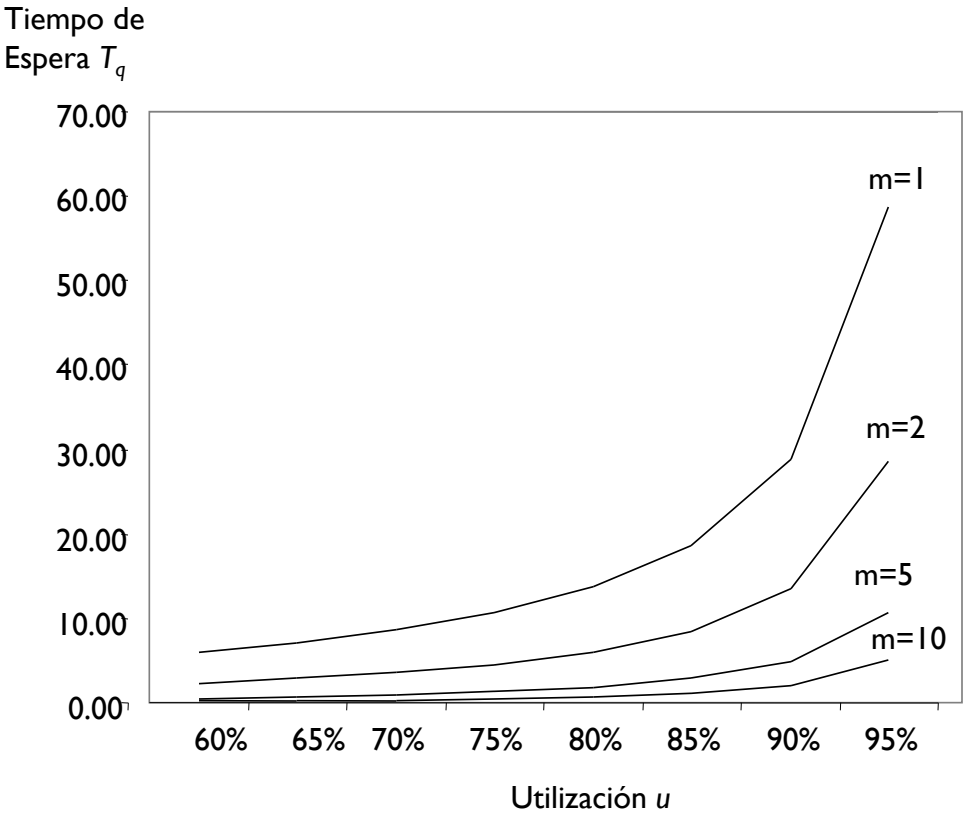
El Poder del Pooling

Recursos Independientes
2x(m=1)



Recursos comunes
(m=2)



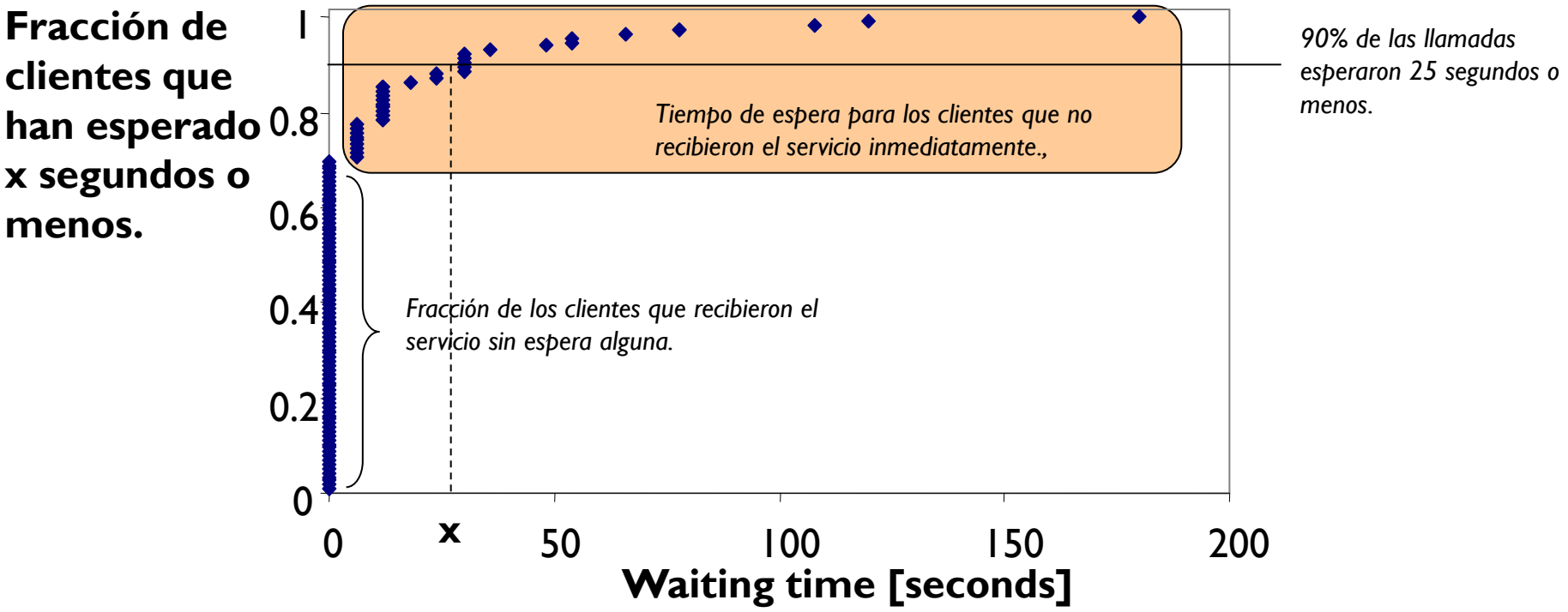


Implicancias:

(+) Utilización balanceada
(+) Capacidad de fondo de seguridad
(+) Economías de escala estadísticas

(-) Cambios / Set-Ups
(-) Menos especialización

Otras medidas de nivel de servicio: Tiempo de Espera Aceptable



- **Tiempo de Espera Aceptable** (AWT)
- **Nivel de Servicio** = Probabilidad { Tiempo de Espera ≤ AWT }

-.

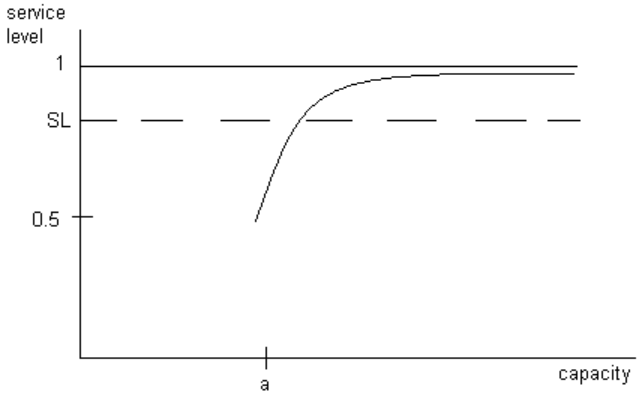
Modelo Erlang Básico: Análisis de la Capacidad.

- $$SL = 1 - P_d(s,a)\, e^{-(s-a)\,AWT\mu} = 1 - P_d(s,\rho)\, e^{-s(1-\rho)\,AWT\mu} \qquad (1)$$

- Asumiendo que un SLA está dado, con AWT y SL acordado (50%). ¿Cuánta capacidad (i.e. cuántos agentes) se necesitan para satisfacer el SLA?
- Usando la aproximación normal

$$SL = 1 - \overline{\Phi}\left(\frac{s-a}{\sqrt{a}}\right) e^{-(s-a)\,AWT\mu}$$

mientras s aumenta de a a ∞. el SL aumenta desde 0.5 a 1.



Análisis de la Capacidad (II)

- Supongamos que se desea cumplir un SLA acordad (dado por un AWT y un SL)
- ¿Cómo crece la capacidad con el volumen de demanda λ

$$s = a + k\sqrt{a}$$

(Fórmula de la raíz del staff)

- (k depende del AWT y de SL)
- La fórmula de la raíz cuadrada del staff muestra las economías de escala y las ventajas en costo de realizar pooling de recursos.

Conclusiones

- La variabilidad es la norma, no la excepción.
  - Se deben medir y comprender las fuentes y tratar de reducirlo.
  - El resto se debe acomodar (por ej. Aumentar la capacidad).
- La variabilidad llevará a tiempos de espera incluso si la utilización es menor al 100%.
- Los modelos de colas son útiles para:
  - Cuantificar los efectos de la variabilidad en la performance del sistema.
  - Analizar distintos escenarios (e.g. reducir los tiempos de servicio promedio, aunar servidores, etc.).

Próxima clase

- Aplicación de los conceptos y las fórmulas.

- Traer Laptop, descargar Excel de U-Cursos.