LA GESTIÓN DE LOS TIEMPOS DE ESPERA

Este documento se cita como

Garcia-Sabater, Jose P. (2020)
La Gestión de las Tiempos de Espera Nota Técnica
RIUNET Repositorio UPV
http://hdl.handle.net/10251/137896

Contenido

1.1 INTRODUCCIÓN

Hay quien ha definido la gestión logística como **la gestión de los tiempos de espera.**

O espera el cliente a que esté el producto, o espera el producto a que esté el cliente, o esperan las máquinas que pueden fabricar el producto inmediatamente para que el cliente disponga de él. O esperan todos, cada uno en su proporción.

Y si no se define apropiadamente quien debe esperar -lo que deba esperar-, todos esperarán a todos porque ese es el modo natural de sincronizarse.

Los tiempos de espera los estudia desde principios del siglo XX una disciplina de la investigación operativa denominada teoría de colas (o de filas o de "tiempos de espera") (Gross *et al.*, 2008). El aparataje matemático necesario para obtener conclusiones es muy importante.

Se denomina "cola, fila o línea" a la cantidad de clientes (pedidos, correos electrónicos, stock…) que están esperando de modo más o menos ordenado a ser atendidos cuando el servidor o servidores queden libres.

En este capítulo no se trata de resumir la teoría de colas (lo que no se puede hacer en el espacio disponible), sino de presentarla como una herramienta potente en el diseño de procesos. Para ello hay que introducir el concepto y así finalmente explicar porqué aparecen las colas. Es relevante en este punto recordar que no es la excesiva carga de trabajo la que provoca la cola sino que es la variabilidad quien la gobierna.

El resto del capítulo se estructura como sigue. En primer lugar, se hace un rápido abordaje a la teoría de colas, se caracterizarán y se presentará la denominada Ley de Little. Antes de seguir se incidirá en la naturaleza estocástica de las colas. A partir de ahí se incorporarán algunas fórmulas que permiten anticipar el tamaño medio de una cola unietapa. Se realizará una aproximación a las redes de colas (que es el modo en el que naturalmente éstas se encuentran) y se finalizará con una aproximación al comportamiento de las colas cuando hay limitación en la capacidad.

1.2 EL ORIGEN DE LAS COLAS

El motivo que hace aparecer las colas no es la falta de capacidad, sino la variabilidad en las llegadas y en el tiempo de operación. Si no sobra capacidad la cola crecerá sin parar. Cuando no sobra suficiente capacidad es que una vez se ha creado la cola, va a tardar tiempo en desaparecer.

"las colas aparecen por la variabilidad y no desaparecen por la saturación"

Si los clientes llegaran de manera regular al sistema, y el tiempo de servicio fuera una constante, no habría colas. Nadie ni nada tendría que esperar. Pero la llegada de los clientes no es regular, ni el tiempo de servicio es siempre el mismo.

Si el tiempo entre dos llegadas consecutivas es aleatorio y el tiempo de servicio sigue otra distribución aleatoria es cuando aparecen las colas.

La cola (el número de clientes esperando a ser atendidos) es estocástica en la medida en la que la tasa de llegada de clientes al sistema sigue una distribución aleatoria y/o el tiempo de servicio a cada uno de los clientes sigue una distribución aleatoria.

1.1 EL SISTEMA MÁS SENCILLO

Para poder entender un sistema de colas hay que empezar por entender la cola más sencilla.

Un conjunto de clientes (pacientes, pedidos, materiales, correos electrónicos…) accede a un servidor (o conjunto de servidores – máquinas, personas-) para recibir un servicio.

Si, cuando el cliente entra, no hay ningún servidor libre, tendrá que esperar a que algún servidor esté libre.

En ese sistema sencillo (una etapa, un tipo de clientes) se pueden definir, en un momento dado tres grupos de clientes la población susceptible de entrar en el sistema, los clientes que esperan a ser atendidos durante un cierto tiempo, y clientes que son atendidos durante un cierto tiempo.

Y el modo en el que entren, esperen y sean atendidos hará que los tiempos de espera para ser atendidos sean mayores o menores, y que el número de clientes esperando sea más grande o más pequeño.

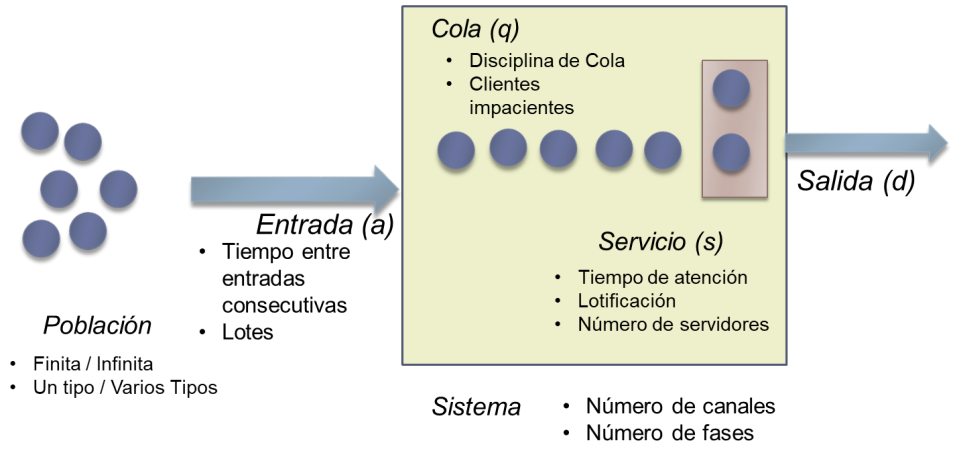

Ilustración 1: Caracterizando una sistema de colas

Conocer el número de clientes que han de esperar permitirá dimensionar *buffers,* informar al cliente, anticipar tiempos de entrega.

La distribución entre llegadas consecutivas, la organización de la cola, el tiempo de servicio y la saturación del sistema influyen en la distribución entre salidas consecutivas. Lo que es especialmente relevante en sistemas industriales.

1.2 PARÁMETROS BÁSICOS PARA DEFINIR UNA COLA UNIETAPA

Los datos necesarios para caracterizar una cola unietapa simple son los que hacen referencia a la llegada de clientes, el tiempo de servicio y el número de servidores que lo prestan.

En teoría de colas convencional se suele utilizar para representar las características básicas del sistema las tasas de llegada (λ : clientes que entran por unidad de tiempo) y la tasa de servicio (μ : clientes que son atendidos por unidad de tiempo).

El ritmo de llegada de los clientes se nombra también como ***tiempo de takt*** *(* $T_a = \frac{1}{\lambda}$*)* y es el tiempo promedio entre llegadas consecutivas (la *a* es de *arrival*)

El ritmo al que se atiende a cada cliente se nombra como ***tiempo de ciclo*** *(*$T_s = \frac{1}{\mu}$*)* *y* es el tiempo promedio de servicio (s minúscula de *service*)

Para poder calcular la saturación del sistema es necesario conocer el número de servidores que van a estar prestando el servicio simultáneamente. Ese parámetro se suele escribir como *S* (aunque muchos libros también utilizan una C).

La combinación de λ , μ y *S* permite calcular la saturación (también llamada utilización o congestión y representada por la letra griega ρ) del sistema que se expresa según la siguiente fórmula. ($\rho = \frac{\lambda}{S\mu} = \frac{T_s}{ST_a}$).

Por definición la saturación efectiva no puede ser superior al 100%. Si eso fuera así la cola crecería indefinidamente puesto que entran más clientes de los que se puede absorber, y de algún modo tendrán que abandonar el sistema sin ser atendidos.

Si el sistema está sobresaturado habrá clientes que se irán (con lo que a la tasa de llegada habría que restarle una tasa de abandono) o habrá que hacer horas extra (con lo que se incrementa la tasa de servicio), o incorporar personal de apoyo (con lo que se incrementa el número de servidores).

La cola esperada crece no linealmente con la saturación. Y a partir del 85% se puede decir que el sistema no será fácilmente predictible en sus tiempos de servicio.

Pero la cola no es el resultado de la sobresaturación sino de la variabilidad (quizá de la sobresaturación "instantánea").

Por eso es importante conocer la distribución estadística que representa los tiempos de servicio (o al menos su media y su coeficiente de variación).

La media ya ha sido identificada como T_a y T_s. El coeficiente de variación, es decir la relación entre la desviación típica y la media, se suele representar como C_a y C_s).

En colas más complejas se puede identificar una población limitada, de uno o varios tipos de clientes, que pueden llegar en lotes, que se pueden organizar con diferentes disciplinas de colas, que pueden abandonar o no el sistema por algún motivo. Pero para analizar todo eso habría que leer otro libro (Gross *et al.*, 2008)

1.3 LEY DE LITTLE

Al describir el comportamiento de un sistema de colas es relevante conocer el tiempo de espera en cola, pero también el tiempo total en el sistema del cliente. Es también relevante conocer el número de clientes que han de esperar, así como el número de clientes que en promedio habrá de esperar.

Quizá el aporte teórico más interesante de la teoría de colas es la denominada Ley de Little que permite relacionar la longitud esperada de la cola con el tiempo de espera promedio y la tasa de entrada de clientes.

Según esta ley la longitud de una cola es proporcional a la tasa de entrada de clientes y al tiempo de estancia esperado en la cola.

$$L_q = \lambda W_q = \frac{W_q}{T_a}$$

Siendo *Wq* el tiempo promedio de estancia en la cola, mientras que *Lq* es el número promedio de clientes en la cola.

Para algunos esto no es una Ley sino una tautología. Para otros es el equivalente a la Ley de Newton en Dirección de Operaciones (Hopp and Spearman, 2001). La ley de Little es válida para cualquier sistema que sea conservativo (es decir que no vaya perdiendo clientes de manera no controlada.)

La ley de Little es intuitiva pero de un modo diferente a como la piensan la mayor parte de los mortales. Cualquier usuario tenderá a pensar que el tiempo de estancia en el sistema es proporcional al tiempo de servicio. Pero esa es "su" percepción individual. La ley de Little aplica a la cola en su conjunto no al cliente individual.

La ley de Little se puede entender con el siguiente simple ejercicio. Si tras acceder a una cola se mide el tiempo que pasa hasta que ser atendido (*Wq*), cuando se eche la vista atrás se verá una cola (de tamaño *Lq*) que será proporcional a la tasa de entrada de clientes en el sistema (λ).

La denominada Ley de Little aplica también para el sistema en su conjunto: el número promedio de clientes en un sistema (L - los que están siendo atendidos y los que no) es proporcional a la tasa de llegada y al tiempo de estancia total en el sistema (W).

$$L = \lambda W = \frac{W}{T_a}$$

De manera intuitiva se puede relacionar Wq con W sin más que restarle a este último el tiempo de servicio ($W = W_q + \frac{1}{\mu} = W_q + T_s$). Por tanto, al estimar alguna de las cuatro variables que identifican el comportamiento promedio de una cola (*L, Lq, W, Wq)* es posible estimar todas las demás.

La ley de Little aplica a sistemas de colas individuales, pero también a sistemas de colas complejos con múltiples etapas.

1.4 COLAS DE UNO Y VARIOS SERVIDORES

A lo largo de casi 100 años, los investigadores en el área han ido obteniendo resultados que permiten describir muchos sistemas de colas: Con diferentes estructuras de llegada y de atención al cliente, con diferente número de servidores y de canales de espera, con reglas de priorización, con diferentes parámetros para caracterizar la llegada y el servicio…

Se puede encontrar una gran variedad de métodos tanto para obtener de modo exacto como para estimar valores. Para poder profundizar en la teoría será relevante conocer la clasificación de *Kendall-Lee*. Dicha clasificación permite (en caso de que fuera necesario acceder a los detalles de la teoría) caracterizar el sistema de colas sobre el que se quiere estudiar su comportamiento.

El sistema más simple de todos es el denominado M/M/1. En ese tipo de sistemas los clientes llegan siguiendo una distribución de Poisson (que básicamente

significa que llegan cuando lo consideran oportuno sin relación entre ellos ni con el paso del tiempo). Un único servidor atiende a los clientes y el tiempo de atención sigue una distribución negativa exponencial (que es una distribución que tiene una desviación típica de 1. (Sorpresivamente la distancia temporal entre dos llegadas consecutivas que sigan un proceso de Poisson, sigue una negativa exponencial, y viceversa).

La probabilidad de que en un sistema M/M/1 tenga n elementos esperando es:

$$P_n = (1-\rho)\rho^n$$

Dicha función establece que existe la posibilidad de que haya valores muy altos, que compensan los momentos en los que no hay nadie en la cola (n=1) o incluso nadie en el sistema (n=0).

Esa función relativamente sencilla permite calcular el tamaño medio esperado de Cola para un sistema M/M/1.

$$L_q = \sum_n n(1-\rho)\rho^n = \frac{\rho^2}{(1-\rho)}$$

Hay que recordar que es el tamaño medio, no el tamaño instantáneo.

En ocasiones el número de servidores es mayor que 1. Se trata de los problemas M/M/c. Muchísimo más complicados en el proceso de cálculo. Complicados pero conocidos, así que es posible calcular la longitud de cola media que dependerá de la saturación.

	num servidores														
saturación	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*15*	*20*	*25*	*30*	*35*
0,5	0,50	0,33	0,24	0,17	0,13	0,10	0,08	0,06	0,05	0,04	0,01	0,00	0,00	0,00	0,00
0,6	0,90	0,68	0,53	0,43	0,35	0,29	0,25	0,21	0,18	0,15	0,07	0,04	0,02	0,01	0,01
0,7	1,63	1,35	1,15	1,00	0,88	0,78	0,70	0,63	0,57	0,52	0,33	0,22	0,15	0,10	0,07
0,8	3,20	2,84	2,59	2,39	2,22	2,07	1,94	1,83	1,73	1,64	1,28	1,02	0,84	0,69	0,58
0,9	8,10	7,67	7,35	7,09	6,86	6,66	6,48	6,31	6,16	6,02	5,42	4,96	4,57	4,24	3,96
0,91	9,20	8,77	8,44	8,17	7,94	7,73	7,54	7,37	7,21	7,06	6,44	5,95	5,54	5,19	4,88
0,92	10,58	10,14	9,81	9,53	9,29	9,08	8,88	8,71	8,54	8,39	7,74	7,22	6,79	6,41	6,08
0,93	12,36	11,91	11,57	11,28	11,04	10,82	10,62	10,44	10,27	10,11	9,43	8,89	8,43	8,02	7,67
0,94	14,73	14,27	13,92	13,63	13,38	13,16	12,95	12,76	12,59	12,42	11,72	11,14	10,66	10,23	9,85
0,95	18,05	17,59	17,23	16,94	16,68	16,45	16,23	16,04	15,86	15,69	14,95	14,35	13,84	13,39	12,98
0,955	20,27	19,80	19,44	19,14	18,88	18,65	18,43	18,23	18,05	17,87	17,13	16,51	15,98	15,52	15,10
0,96	23,04	22,57	22,21	21,91	21,64	21,40	21,18	20,98	20,79	20,62	19,85	19,23	18,69	18,21	17,77
0,965	26,61	26,13	25,77	25,46	25,19	24,95	24,73	24,53	24,34	24,16	23,38	22,73	22,18	21,68	21,24
0,97	31,36	30,89	30,52	30,21	29,94	29,69	29,47	29,26	29,07	28,88	28,09	27,43	26,86	26,35	25,89
0,975	38,03	37,54	37,17	36,86	36,58	36,34	36,11	35,90	35,70	35,51	34,71	34,03	33,45	32,93	32,45
0,98	48,02	47,53	47,16	46,84	46,57	46,31	46,08	45,87	45,67	45,48	44,66	43,97	43,37	42,83	42,34
0,985	64,68	64,19	63,81	63,49	63,21	62,96	62,73	62,51	62,30	62,11	61,27	60,57	59,96	59,41	58,90
0,99	98,01	97,52	97,14	96,81	96,53	96,27	96,03	95,81	95,61	95,41	94,56	93,84	93,21	92,64	92,12
0,995	198,01	197,51	197,12	196,80	196,51	196,25	196,01	195,78	195,57	195,38	194,51	193,77	193,13	192,55	192,01

Tabla 1: Valor de Lq de un sistema MMc en función de la saturación del sistema y del número de servidores

A partir del valor de Lq es posible calcular según se ha visto antes L, W y Wq. Pero es muy importante recordar que son valores medios.

1.5 LA NATURALEZA ESTOCÁSTICA DE LA COLA

Sería sospechoso que el número de clientes en una cola fuera siempre el mismo. La variabilidad en los tiempos de llegada y los tiempos de atención generan variabilidad en la cola. Y es interesante destacar una obviedad estadística: el tiempo promedio de estancia no es el tiempo de estancia moda ni el tiempo de estancia promedio.

Es relevante recordar que una observación particular no configura la realidad de la cola. Cada cliente de una cola tendrá una observación distinta. Los clientes observan sólo cuando están ellos y sólo miran lo que tienen delante y el tiempo que ellos han estado.

La longitud media de cola no es lo que la mayor parte de los clientes van a observar, puesto que la cola no se distribuye según una distribución de tipo normal.

La longitud de la cola está limitada por abajo en 0, no estando limitado por arriba.

Utilizando la fórmula que permite derivar la probabilidad de que haya N clientes en una cola en la que la entrada siga una distribución de Poisson y el tiempo de servicio una negativa exponencial (un caso particular pero muy habitual) se puede representar el gráfico de la Ilustración 138: P(n) de un sistema MM3 al 83,3%

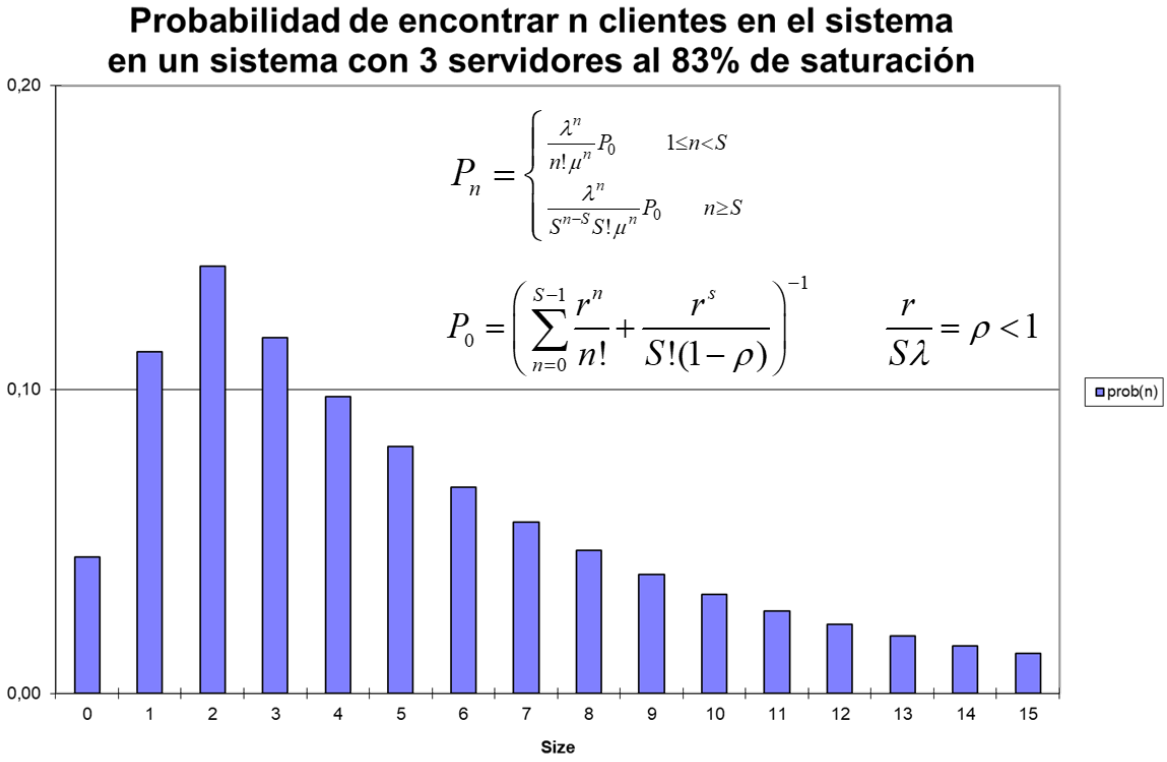

Ilustración 2: P(n) de un sistema MM3 al 83,3%

Como se puede observar, pese a que el nivel de saturación no es muy elevado (83%) existe la posibilidad de que haya muchos clientes esperando porque estadísticamente pueden llegar muy juntos y coincidir con tiempos de servicio elevados. Y cuando una cola al 83% crece mucho, el sistema necesita atender a 6 clientes para reducir en uno la cola.

1.6 APROXIMANDO EL TAMAÑO DE LA COLA A LOS SISTEMAS GENERALES (G/G/S)

Se propone a continuación una aproximación para poder entender y aplicar más rápidamente el concepto. La propuesta extiende las soluciones analíticas complejas a una fórmula aproximada que aplica a niveles de saturación altos (los más útiles) y hace más evidente la utilidad del concepto.

Para poder interpretar la fórmula que se propone es necesario explicitar dos nuevos parámetros. Son el coeficiente de variación al cuadrado de los tiempos entre llegadas consecutivas (C_a^2) y el coeficiente de variación al cuadrado de los tiempos de servicio (C_s^2). El coeficiente de variación es la relación entre la desviación típica y el valor promedio, en este caso la desviación típica de los tiempos entre llegadas consecutivas y la media de tiempo entre llegadas

consecutivas (que se obtiene al calcular el tiempo total dividido entre el número de llegadas al sistema).

Valores elevados de coeficiente de variación indican un sistema con una alta variabilidad. Valores reducidos de coeficiente de variación indican un sistema que ha sido puesto bajo control donde las llegadas siguen un patrón regular.

Por poner una referencia, un sistema en el que los clientes llegaran de uno en uno, en cualquier momento y sin relación entre ellos (llegadas según Poisson, tiempo entre llegadas siguiendo una negativa exponencial) tendrían un coeficiente de variación de 1. Mientras que un sistema del que se pueda programar las llegadas se puede conseguir un coeficiente de variación entre llegadas consecutivas nulo.

Se conoce como la aproximación Allen-Cunneen y tiene la siguiente expresión.

$$Lq \approx \left(\frac{C_a^2 + C_s^2}{2} \right) \frac{\rho^{\sqrt{2S+2}}}{(1-\rho)}$$

Esta fórmula es una aproximación suficiente para sistemas saturados. Es decir, aproxima mal para situaciones de saturación bajas. Pero permite entender de manera el comportamiento no lineal de un sistema de colas en el entorno que interesa, el de alta saturación.

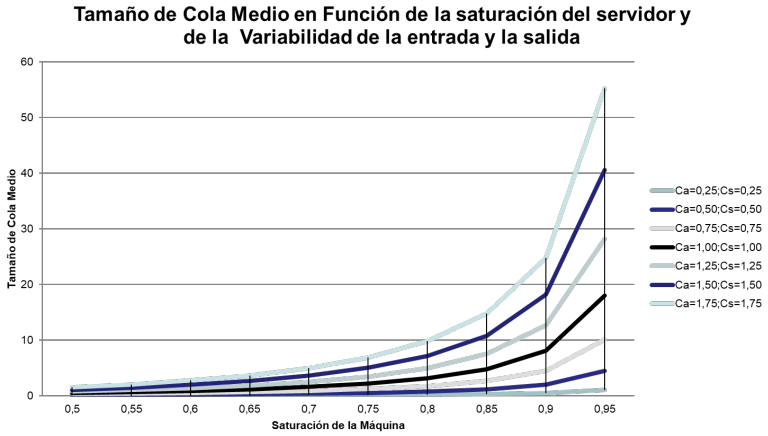

Ilustración 3: Tamaño de Cola Medio en función de la saturación del servidor y de la variabilidad a la entrada y a la salida

Cuando la saturación es elevada (cercana a 1) el tamaño de cola crece de manera no lineal. La fórmula lo expresa al dividir por (1-ρ).

Si la saturación es baja el tamaño de cola será bajo, tanto más bajo cuantos más servidores en paralelo haya. Pero el verdadero modo de reducir el tamaño de la

cola es reducir la variabilidad de los tiempos de servicio y los tiempos entre entradas consecutivas.

1.7 REDUCIENDO O LIMITANDO EL TAMAÑO DE LA COLA

Analizando la fórmula de Allen-Cunneen se pueden intuir cuatro modos de reducir el tamaño esperado de una cola:

1. Reducir la variabilidad del tiempo entre llegadas consecutivas. En ocasiones es posible programar las llegadas, o al menos espaciar algunas (sistemas de cita previa, por ejemplo).
2. Reducir la variabilidad del tiempo de servicio se puede lograr con la estandarización de las operaciones, generación de programas de mantenimiento (TPM y similares), y técnicas como el SMED.
3. Reducir la saturación del sistema (desincentivando la llegada de clientes o reduciendo el tiempo de servicio)
4. Incrementar el número de servidores en paralelo. Incrementar el número de servidores es la solución inmediata, y también la más costosa. Una alternativa no tan cara es agregar diferentes sistemas para que se comporten como uno solo. La agregación reduce *Lq*, pero fundamentalmente reduce *Wq* (pues reduce *Ta*).
5. Limitar la capacidad de la cola. Es decir, no se aceptan más clientes a partir de una determinada cantidad esperando. El número de clientes que se pierde sería la tasa de llegada por la probabilidad de que el sistema esté lleno. A cambio se garantiza a los clientes a los que se les atiende en un tiempo razonable. En teoría de colas se conoce a este problema como problema de cola limitada o G/G/S/K. Siendo K la capacidad del sistema incluyendo a los s servidores.

En las empresas de fabricación este esquema es muy utilizado porque permite además de controlar la cantidad de stock en una determinada sección poder garantizar tiempos de entrega y tiempos de tránsito adecuados. Se trata de controlar el stock delante de una máquina (reduciendo el número de soportes para la unidad de carga o limitando el espacio disponible). En ese caso alcanzar el límite del almacén provoca un bloqueo en las máquinas anteriores que, en buena lógica debieran parar de producir, subiendo el bloqueo “aguas arriba”.

Una variante del sistema es aquella en la que se ponen carteles anunciando el tiempo de espera si la cola está en un determinado punto, animando a que un porcentaje de los posibles nuevos clientes no se incorporen porque “ya pasarán más tarde”. Lo que equivale a que el cliente impaciente abandone el sistema (pero de manera más elegante advirtiéndoselo).

Otra variante es incorporar una etapa previa a la cola (un “portero”) que impide que entre más gente en el servicio principal que se quiere proteger. Una variante más

sofisticada de ésta haría que el portero se dedicara a "diferir" a los clientes que llegan por la vía de ofrecerles una cita en otro momento.

1.8 REDES DE COLAS

Las colas, como cualquier otro tipo de desgracia, nunca vienen solas. E incluso como se acaba de ver en el apartado anterior es bueno que sea así).

El análisis de un sistema de colas en red es sustancialmente más complejo que una red única.

En cada una de las colas hay dos tipos de entradas, las que viene de fuera (γ) y las que vienen como salida de otros nodos. Los flujos de entrada se acumulan y conforman la entrada a cada nodo.

De cada nodo salen clientes que pueden ir fuera del sistema o a otros nodos (incluyendo el mismo nodo del que acaban de salir). Esos movimientos se representan mediante una matriz de transición $r_{i,j}$ que representan el porcentaje de clientes que salen que van al nodo j de entre los que salen del nodo i.

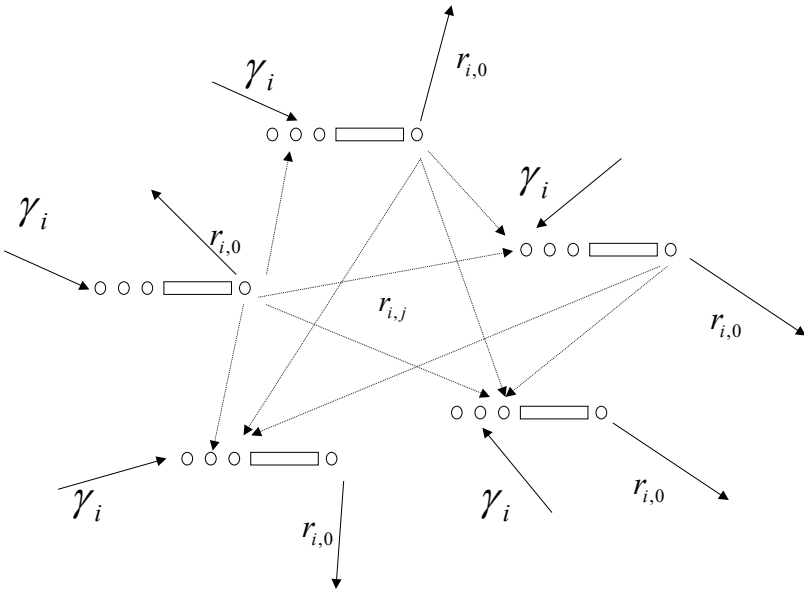

Ilustración 4: Red de Colas

Es interesante nota que la ley de Little aplica para las redes de colas tanto como para los sistemas individuales.

$$\sum_i L_i = W \sum_i \gamma_i$$

En la literatura científica las redes más sencillas son aquellas que representan sistemas en serie. Un poco más complejas son las Redes de Jackson en las que todas las etapas son redes M/M/S. Sustancialmente más complejas son las redes de colas cerradas, donde no hay clientes externos que entren en el sistema. En estos apuntes se va a trabajar con una aproximación muy aproximada en la que todas las redes son de tipo general.

1.8.1 LO QUE SALE DE UNA ETAPA

Para analizar el comportamiento de las colas cuando están en red es relevante entender la salida de un sistema de cola. Porque que la salida de una etapa sea la entrada de lo siguiente es lo habitual en cualquier proceso (que por ello es proceso, porque hay diferentes etapas).

Del mismo modo que en párrafos anteriores se ha indicado que los datos relevantes de la entrada y el servicio es el tiempo medio entre llegadas o entre atenciones, y el coeficiente de variación de esos tiempos, describir la distribución de la salida sería describir la distribución de tiempos entre salidas consecutivas. Del mismo modo que se ha nombrado con el subíndice *a* al subsistema llegadas, y con el subíndice *s* al subsistema servicio se denomina con el subíndice *d* al subsistema salida.

Como parece evidente, salvo que el servicio sea de destrucción, la tasa de entrada debe ser igual a la tasa de salida. Si entra λ sale λ. Y por tanto $T_d = T_a = \frac{1}{\lambda}$

La diferencia entre la entrada y la salida se aprecia en la variabilidad entre las salidas consecutivas. La fórmula que permite aproximar el coeficiente de variación de las salidas es la siguiente.

$$C_d^2 \approx \left(1-\rho^2\right) C_a^2 + \rho^2 \frac{C_s^2 + \sqrt{S} - 1}{\sqrt{S}}$$

Cuando el sistema está poco saturado la salida se parecerá a la entrada. Como parece natural es servidor o servidores, al no haber cola (resultado de la poca saturación), se limitarán a dejar pasar lo que entra.

Si el sistema tiene una saturación elevada próxima la salida estará condicionada por el servicio de esa cola.

Este hecho permite controlar la variabilidad en los sistemas si se utiliza adecuadamente. Permite eliminar la variabilidad de la entrada en aquellos subsistemas donde la cola tiene un efecto nefasto en el servicio. Concentrando el

problema en un único sitio, y permitiendo trabajar los tiempos de espera dónde interesa a la empresa.

Si, por ejemplo, un restaurante establece un sistema de recepción antes de entrar en la sala, los clientes esperarán "de pie" a que les asignen mesa. Cuando al cliente se le asigne una mesa el resto de servicios (camareros y cocina) podrán trabajar de manera regular y su sincronización será posible. Y al haber concentrado el tiempo de espera en un único lugar, se podrá trabajar ese ámbito, convirtiendo la experiencia en algo positivo. Este concepto aplica también a fábricas, hospitales, aduanas …

1.8.1 CONFLUENCIAS Y BIFURCACIONES

Las redes de colas tienen como característica que los flujos confluyen o se bifurcan (no serían una red en otro caso). Así que parece razonable analizar el comportamiento de los tiempos entre sucesos consecutivos en estos casos.

Si una fila se alimenta a partir de varias filas que confluyen, es conveniente saber que los coeficientes de variación al cuadrado se agregan de manera ponderada a la cantidad de población que entra desde cada flujo.

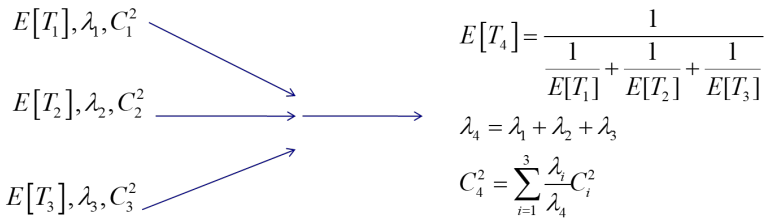

Ilustración 5: Confluencia de Flujos

En ocasiones, al recibir el servicio, las filas se dividen. Se propone un modo para estimar cual sería el coeficiente de variación al cuadrado de los tiempos de paso para cada una de las filas.

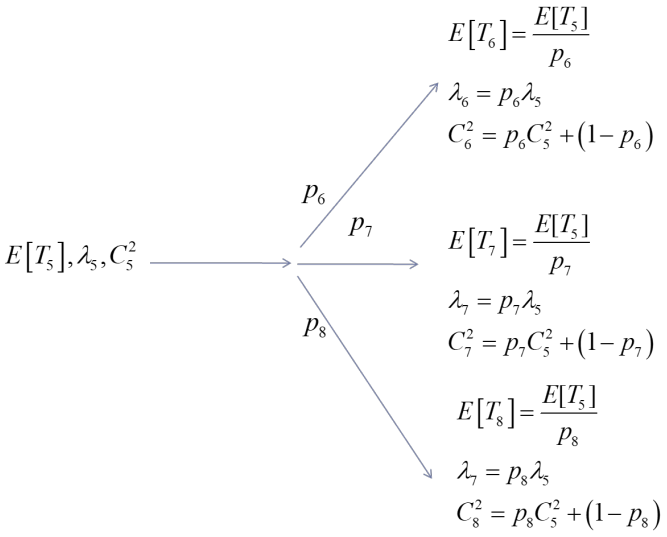

Ilustración 6: Bifurcaciones de Flujos

Pese a toda la complejidad que estas fórmulas sugieren, es interesante conocer que la Ley de Little es válida para un sistema formado por un conjunto de subsistemas de colas conectados entre ellos.

Representando la red con tasas de entrada a cualquiera de los sistemas desde el exterior representadas por γ (no se representa como λ porque esa letra se guarda para incluir también los clientes que vienen desde otros lugares de la red, la cantidad de clientes en el sistema será proporcional al número de clientes que entran desde el exterior y el tiempo de espera global.

$$\sum_i L_i = W \sum_i \gamma_i$$

1.9 ¿DE VERDAD HE DE ENTENDER ESTAS FÓRMULAS?

El atento, y por ello caro, lector se estará preguntando continuamente ¿*ytoestopaqué*?.

Con el ánimo de animar a entender mejor la teoría de colas y sus implicaciones se expone a continuación un ejemplo que permitiría visualizar la utilidad de entender bien el concepto.

Una empresa tiene un servicio de atención al cliente con un equipo de "asistentes comerciales" que trabajan un número de horas al día. El proceso actual es el siguiente:

1. El cliente accede a la aplicación informática y solicita que le atiendan, para garantizar la seguridad del proceso, el cliente es asignado a uno de los "asistentes comerciales" que forman el equipo operativo de la unidad.

2. El cliente espera "pacientemente" a que le llame el "asistente comercial".

3. Algunos de los pacientes clientes se quejan de los largos tiempos de espera en el contestador automático que han puesto para medir el tiempo de servicio con la inútil esperanza de que alguien mejore el servicio.

4. Se reúnen los del departamento de calidad con los del servicio al cliente y le dicen que no pueden hacer nada que están muy saturados.

5. Gerencia dice que no va a contratar a nadie más porque la saturación no llega al 90%.

Un análisis detallado indica que la llegada de los clientes sigue una distribución de Poisson y que el tiempo de servicio tiene una alta variabilidad.

Algunas opciones que se abren a la luz de la teoría:

1. Reducir la variabilidad de los procesos quizá especializando a los asistentes, quizá incorporando soporte informático o de otro tipo
2. Asignar ventanas de entrada a los clientes, quizá modificando sus patrones de solicitud.
3. Cambiar el modo y momento en el que se asigna al asistente.
4. Dividir el proceso en etapas consecutivas de igual duración.

La teoría de colas puede dar luz sobre el origen del problema y sobre la pertinencia de las soluciones. En función de la combinación de números (tasas de entrada y de salida, variabilidad, capacidad de dividir el proceso…) la solución adecuada es una u otra. Y los números se relacionan entre sí no de una manera directa. Incluso en el caso de que no se reduzca la saturación (o incluso aunque ésta crezca añadiendo alguna tarea adicional) el tamaño de la cola puede que se reduzca. E incluso aumentando el tamaño de cola podría reducirse el tiempo de estancia, porque son magnitudes diferentes (aunque muy relacionadas).

BIBLIOGRAFÍA

Gross, D. *et al.* (2008) *Fundamentals of queueing theory*. Wiley.

Hopp, W. J. and Spearman, M. L. (2001) *Factory physics foundations of manufacturing management Wallace J. Hopp, Mark L. Spearman*.