ESCOLA TÈCNICA SUPERIOR D'ENGINYERIA DE TELECOMUNICACIÓ DE BARCELONA

TRÁFICO DE TELEFONÍA MÓVIL: CARACTERIZACIÓN E IMPLICACIONES DEL TIEMPO DE OCUPACIÓN DEL CANAL

Autor: Francisco Barceló Arroyo
Director: Josep Paradells Aspas

CAPÍTULO 1

Modelos de teletráfico en sistemas de telefonía móvil

En este capítulo se introducen conceptos relativos al teletráfico, junto a los modelos habitualmente utilizados para el dimensionado y evaluación de sistemas de telefonía móvil. A lo largo de esta tesis se consideran como sistemas de telefonía móvil aquellos cuya finalidad primordial, aunque no exclusiva, es la de proporcionar servicios de voz. De este modo contemplamos la posibilidad de que la voz coexista en ellos junto a otros servicios, principalmente de datos, proporcionando a la red un cierto valor añadido. Los servicios de voz representan en la actualidad entre el 80% del tráfico en redes avanzadas de telecomunicación [MAZ96a]. Previsiones a largo plazo otorgan una participación del 66% [BAR96e] a los servicios que no toleran retardos y que hoy requieren conmutación de circuitos (telefonía, videotelefonía y videoconferencia). El uso extendido de los modelos de colas M/M/*s*/*s* o "Blocked-Calls-Cleared" y M/M/*s* o "Blocked-Calls-Delayed" aplicado a este tipo de sistemas, que se resuelven respectivamente mediante las conocidas ecuaciones de Erlang-B y Erlang-C, es sin duda una herencia del dimensionado de sistemas de telefonía fija, tal vez innecesaria, propiciada por la ausencia de soluciones para otros modelos más cercanos a la realidad de los sistemas de telefonía móvil. En capítulos sucesivos se introducen modelos en los que no se cumple la hipótesis de distribución exponencial (M) del tiempo de servicio, y para los cuales no existen soluciones analíticas exactas, sino únicamente resultados aproximados.

1.1 Los tres niveles de tráfico de un diálogo

En sistemas de telefonía pueden establecerse tres niveles de tráfico de generación de señal vocal tal como se ilustra en la figura 1.1. El nivel superior es el de

llamada (también conocido como de *conversación* o *mensaje*) que está activo durante toda la fase de diálogo. El ejemplo más claro lo constituye la llamada telefónica en full-duplex a través de la RTP (Red Telefónica Pública) para la que ocupamos un canal a lo largo de toda la conversación, aunque al menos el 50% del tiempo estamos callados si es que deseamos entender lo que nos está diciendo nuestro interlocutor. Sin embargo y dado que disponemos de canal, sería posible aunque poco práctico, que ambos interlocutores hablaran durante todo el tiempo.

El segundo nivel es el de *transmisión*, para el cual solo se asume activo el canal del interlocutor que efectivamente está hablando, siendo la media de la actividad a este nivel inferior al 50% de la actividad a nivel de llamada, debido a que durante parte del tiempo de la llamada el sistema está ocupado con señalización. La asignación de recursos en base a la transmisión puede mejorar obviamente la eficiencia de los mismos, aunque dicha mejora dependerá de la calidad que se exija a la red, que puede ser diferente para ambos niveles de asignación (ver sección 1.6). La asignación de recursos a nivel de transmisión solamente es posible en sistemas en los cuales el usuario es consciente de la transmisión en sí misma, por ejemplo apretando un botón en sistemas PMR ("Private Mobile Radio"), ya que de hecho el nivel de transmisión no es detectable fácilmente de forma automática como lo es el nivel de ráfaga. Por esta razón este nivel no aparece en algunos modelos jerárquicos de tráfico propuestos pensando en redes fijas como [FIL91, KUE96] en los que de forma implícita se supone al usuario libre de cualquier trabajo relacionado con el esquema de la comunicación.

No durante todo el intervalo de transmisión estamos generando señal vocal, debido a que necesitamos respirar, pensar y realizar pausas para una mejor expresión. La parte efectivamente activa de una transmisión constituye el nivel más bajo de tráfico telefónico y es denominada *ráfaga* ("talkspurt"). Los sistemas basados en interpolación de palabra solo ocupan el canal durante el tiempo de ráfaga, obteniendo así una mayor eficiencia de los recursos siempre dependiente de la calidad exigida a cada nivel. Para ello precisan de un DAS (Detector de Actividad Silencio). Los sistemas de cables transoceánicos utilizan este principio de conmutación a nivel de ráfaga desde la década de los 60: sistemas TASI ("Time Assigned Speech Interpolation") o su versión digital DSI ("Digital Speech Interpolation") [CAM87]. Está previsto que los sistemas de telefonía móvil de tercera generación también utilicen este principio explotado por protocolos de acceso al medio como el PRMA ("Packet Reservation Multiple Access") [GOO90] para mejorar la eficiencia espectral. También la transmisión de voz sobre redes de banda ancha está prevista a nivel de ráfaga [HUI90]. Un estudio analítico de los sistemas basados en

interpolación de palabra tales como el TASI o el TASI con almacenamientos de retardo intermedio lo podemos encontrar en [DES85]; en dicho trabajo se utiliza como punto de partida el modelo de Poisson M/M/∞.

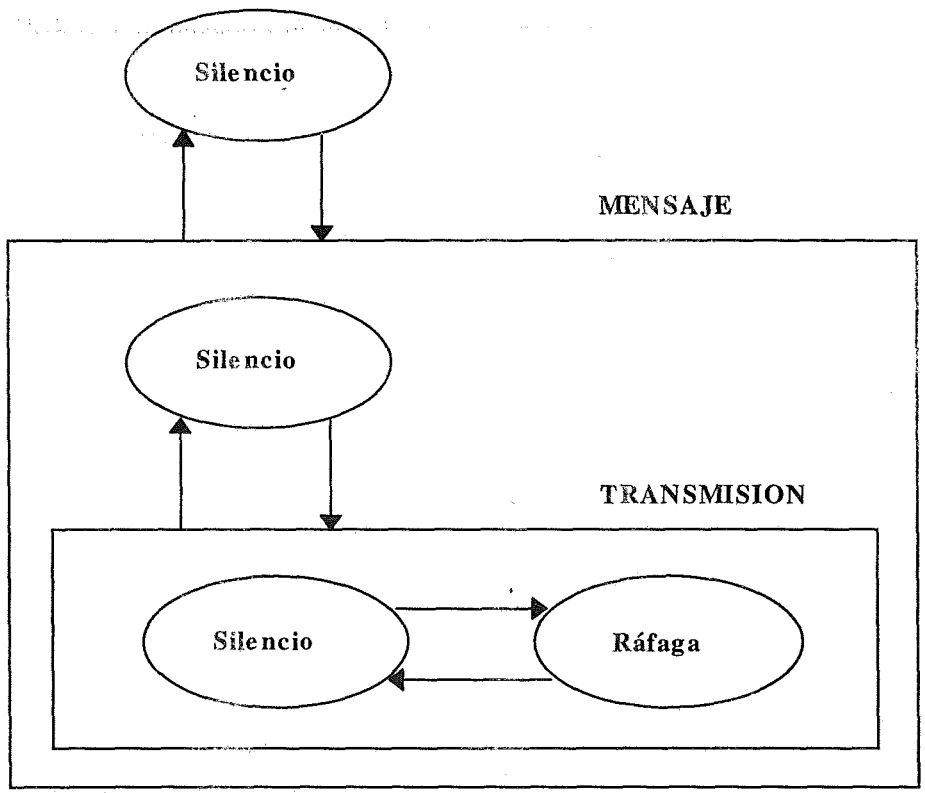

Figura 1.1 Los tres niveles jerárquicos del tráfico vocal.

1.2 Modelos de tráfico para los tres niveles

Los modelos de tráfico más comúnmente utilizados en la literatura para los tres niveles son los siguientes:

- Para el nivel de llamada suele asumirse que la duración de la misma está distribuida exponencialmente. Bolotin refiere en [BOL94b] hasta ocho ponencias en un solo congreso (13^{th} International Teletraffic Congress) que presuponen esta distribución de la duración del mensaje. Suele asumirse también que las llegadas de llamadas son de Poisson. Podemos citar [HON86, GUE87, STE92, ZON95, GAV96] como muestra de referencias incluidas en esta tesis que asumen ambas distribuciones para la duración de la llamada en sistemas de telefonía móvil pública. Las referencias [DOG86, HAS87,

CHR91, HOA91, SIN94] también asumen las distribuciones citadas para sistemas de telefonía móvil de grupo cerrado. Si las llegadas de llamadas siguen un proceso de Poisson el intervalo de tiempo entre llegadas sucesivas sigue una distribución exponencial negativa. Si la actividad es baja (en la RTP pueden considerarse unos 58 mili-Erlang por usuario medio en la hora cargada [BAR96e]) el tráfico generado por un usuario puede considerarse de Poisson, ya que el tiempo entre llegadas de mensajes será prácticamente igual al intervalo de silencio distribuido exponencialmente.

- Para el nivel de transmisión puede asumirse que dentro de un mensaje tanto el periodo de actividad como el de silencio están distribuidos exponencialmente. De hecho si no se incluye la señalización en el modelo (tarea compleja y que proporcionaría resultados excesivamente dependientes del sistema concreto) el tiempo de transmisión de un interlocutor es el de silencio del otro. Así, ambos tiempos, transmisión y silencio entre transmisiones sucesivas, deben seguir la misma distribución estadística.

- Para el nivel de ráfaga suele tomarse el mismo comportamiento que para los dos anteriores, es decir ambos actividad y silencio distribuidos exponencialmente dentro del mensaje [GOO90]. Existe sin embargo la recomendación P-84 de la ITU-T basada en [LEE86a] que establece distribuciones hipergeométricas tanto para los periodos de actividad como para los de silencio sobre una base discreta de tiempo. Estas distribuciones pasan a ser hiperexponenciales sobre una base de tiempos continua (ver sección 4.7).

Si se acepta la hipótesis de que todas las distribuciones implicadas en la conversación son exponenciales, los modelos de fuente a nivel de transmisión y de ráfaga serán IPP ("Interrupted Poisson Process" o proceso de Poisson interrumpido), que es a su vez un caso particular de SPP ("Switched Poisson Process") [KUC73]. Este modelo consiste en un proceso de Poisson imbricado dentro de otro, es decir, tráfico de Poisson únicamente dentro de los periodos de actividad de otro proceso de Poisson de nivel superior. La distribución de los periodos de actividad es exponencial negativa con la tasa de la ráfaga, y la de los periodos de silencio es hiperexponencial-2 [GIR90]. Este modelo de fuente se ilustra en la figura 1.2 y es habitual su utilización en el modelado de sistemas MTA [KUE96].

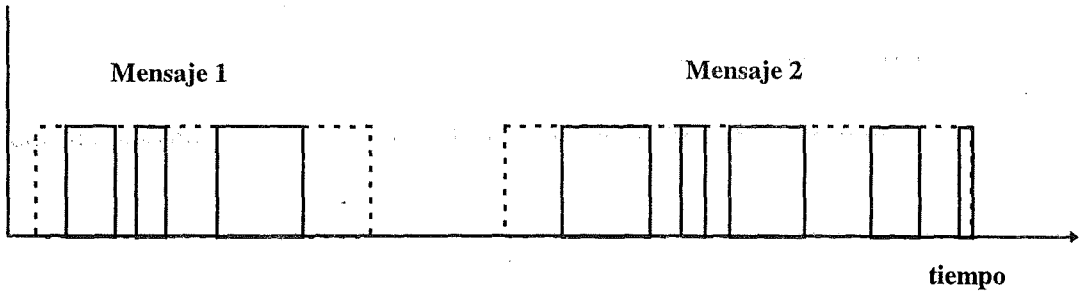

Figura 1.2. Proceso de Poisson interrumpido (IPP).

Estos modelos que resultan en muchos casos ajustados a la realidad del sistema que se pretende evaluar, son en otras ocasiones excesivamente simplistas. El abuso de los mismos se debe, tal como se ha dicho en la introducción, a razones de herencia histórica, y también es debido en parte a la facilidad con la que se obtienen soluciones analíticas exactas a partir de las hipótesis realizadas. En los capítulos 2 y 4 se analizan con más detalle los modelos de voz en tráfico telefónico de redes móviles y se opta por modelos más ajustados a la realidad procedentes de la literatura y de medidas realizadas en sistemas de telefonía móvil.

1.3 Sistemas PMR troncales frente a sistemas monocanal

Los sistemas PMR son sistemas de telefonía móvil de grupo cerrado en los que los clientes del sistema son flotas de vehículos. Tradicionalmente se han venido utilizando sistemas de radio monocanal para dar servicio a este tipo de llamadas, en general de contenido profesional. El sistema troncal se basa en la compartición de los canales de modo dinámico entre un grupo de flotas, de modo que la agrupación de tráficos y de canales conduce a un mejor GoS ("Grade of Service" o grado de servicio) en términos de retardo medio de acceso o probabilidad de demora (ver sección 1.4). El argumento de un mejor GoS para un número de canales determinado se puede invertir: si fijamos un GoS satisfactorio necesitamos en general menos canales para obtenerlo con un sistema troncal que con un conjunto formado por el mismo número de sistemas monocanal. De este modo se obtiene un ahorro de ancho de banda espectral que puede ser importante. Es obvio por otra parte el ahorro que se produce en infraestructura en general debido a la compartición de equipos entre todas las flotas.

En el caso de que un sistema PMR troncal sea gestionado por un operador público es conocido como PAMR ("Public Access Mobile Radio"). En tal caso el operador suele ofrecer su servicio a flotas de vehículos con una tarifa plana: a coste fijo independiente de la cantidad y duración de las llamadas. De cara a mejorar los parámetros del GoS dinaminazando el uso de los recursos el sistema suele imponer una limitación de la duración de la llamada. Dicha limitación es fácilmente aceptada por los cliente empresarios de las flotas, que no están excesivamente interesados en que las conversaciones de los trabajadores puedan alargarse más allá de lo estrictamente necesario. Además suele establecerse un sistema de prioridades de acceso que mejora mucho la demora de acceso de unos pocos a costa de empeorar algo la demora del resto: pensemos en un servicio de ambulancias frente a un transportista comercial. Algunos organismos como ayuntamientos o comunidades autónomas pueden poseer también sistemas troncales PMR corporativos, no dedicados a una explotación comercial y de acceso restringido a las flotas pertenecientes a la entidad.

La tecnología actualmente utilizada para los sistemas PMR troncales es analógica y el standard MPT1327 ha sido ampliamente aceptado [MAZ96b]. En Europa está prevista en breve la aparición del standard TETRA (Trans European Trunked RAdio) [CLI92, ETR96] del ETSI (European Telecommunications Standards Institute) de tecnología digital y basado en la segunda generación de sistemas digitales de telefonía móvil.

La más avanzada tecnología de los sistemas troncales permite establecer entre otras las siguientes diferencias en las clases de llamadas que se producen con respecto a los sistemas PMR convencionales monocanal:

- El modo *"canal abierto"*, consistente en la posibilidad de hablar y escuchar todos con todos en un determinado subgrupo de terminales, es un servicio o canal lógico y no un radiocanal asignado estáticamente al grupo en cuestión.

- La comunicación *full-duplex* es posible y constituye un simple problema de gestión de recursos. Aunque desaconsejable de cara a maximizar la eficiencia espectral, este servicio resulta interesante en la interconexión con redes como la RTP (Red Telefónica Pública) ya que resulta incómodo el hecho de que uno solo de los interlocutores deba utilizar un esquema de "cambio". También es útil para mejorar la percepción de calidad de algunos usuarios no habituados a la disciplina que impone la comunicación vocal en *half-duplex*.

- Las *llamadas de datos* están presentes con una frecuencia muy superior a la que se produce en sistemas monocanal donde el servicio de datos es marginal. Los mensajes cortos constituyen una servicio de especial interés destinado a suplir en parte el servicio de radiobúsqueda ("Pagers").

- También es habitual el establecimiento de sistemas de *prioridad* que posibilitan el acceso rápido de algunas llamadas.

1.4 Modelo para sistemas PMR con asignación por mensaje

Los sistemas PMR troncales con asignación por mensaje disponen de un conjunto de s servidores o canales que se asignan a los usuarios durante toda la duración de la llamada o mensaje. Al producirse el acceso a la red a nivel de llamada, los parámetros más relevantes del GoS son los siguientes:

- *Probabilidad de demora*: es la probabilidad de que una llamada sea demorada cuando intenta acceder a la red. Es por tanto la probabilidad de que no exista un canal disponible en el momento en que la llamada pretende acceder a la red.

- *Retardo medio de acceso*: es el tiempo que tarda en media una llamada en acceder a la red. El cálculo de esta media considera tanto las llamadas demoradas como las no demoradas. Este es un parámetro que el usuario percibe de un modo muy intuitivo: saber que la probabilidad de demora en el acceso a una red es del 5% nos dice poco (en el 5% de los intentos de acceso sufriremos un retardo grande o pequeño) mientras que saber que el tiempo medio de acceso es de 5 segundos nos da una idea bastante clara de la calidad del sistema. Este es un parámetro ampliamente utilizado en estudios de teletráfico de redes móviles [HES81, AVE89, GRI91, HER93].

- *Porcentaje de llamadas demoradas*: es el porcentaje de llamadas que sufren una demora mayor que una cierta cota de tiempo. La probabilidad de demora coincide por tanto con el "percentil 0" o probabilidad de que una llamada sea demorada un tiempo cualquiera. Esta valor del GoS puede ser fijado en términos absolutos, aunque la ITU-R [CCI90] sugiere como GoS el porcentaje de llamadas que se demoran un tiempo mayor que la duración media de una conversación, es decir en términos relativos al tiempo medio de ocupación del canal [HER93].

La evaluación y dimensionado de sistemas PMR troncales suele realizarse en base a un modelo de cola infinita M/M/*s* ("Blocked-Calls-Delayed") [HES81, HAS87, HOA91, HER93, SIN94, MAZ96] para el que existe una solución analítica sencilla y los parámetros del GoS pueden obtenerse fácilmente de forma exacta. El modelo es el representado en la figura 1.3. y las hipótesis que dicho modelo asume se discuten a continuación. Algunos valores típicos de teletráfico (tráfico ofrecido y GoS) en sistemas troncales pueden encontrarse en [HER93, MAZ96].

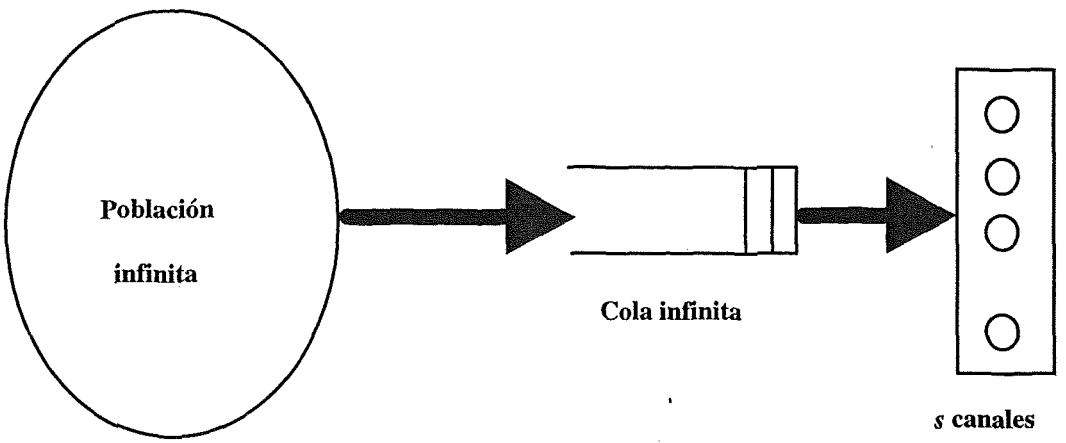

Figura 1.3 Modelo de sistemas PMR troncales con población infinita.

La hipótesis de *población infinita* implica que la población está constituida por muchos usuarios y que las demandas de servicio son independientes entre sí. El concepto de "muchos" es siempre relativo y en este caso debe ser un número alto en comparación al número de canales disponibles. El hecho de que el tráfico generado por terminal sea bajo refuerza esta hipótesis. Es habitual considerar un tráfico de 5,5 mili-Erlang por terminal, resultante de una llamada de 20 segundos cada hora [HER93, MAZ96]. Esta situación debe encontrarse en sistemas PAMR ya que si el número de usuarios es bajo son inviables económicamente. Por otra parte el hecho de que las flotas clientes de los sistemas PAMR sean diversas refuerza la hipótesis de independencia.

El modelo M/M/*s* implica que el tiempo de servicio, en este caso la duración del mensaje, está distribuido exponencialmente. Aquí sí que se está realizando una hipótesis simplista ya que está comprobado (ver capítulo 2) que en sistemas PAMR que transportan solamente información vocal, la distribución de la duración del mensaje es hipoexponencial (a ello contribuye la limitación de la duración de la llamada habitual en estos sistemas), con lo que los resultados del modelo M/M/*s*

llevarán a predecir un GoS peor que el que se va encontrar luego en el sistema real. De modo equivalente se puede decir que la aplicación del modelo M/M/*s* conduce a la sobredimensión del sistema. Por otra parte si el sistema analizado transporta datos además de voz, puede llegarse fácilmente a distribuciones hiperexponenciales de la ocupación del canal, dada la dispersión en la duración de ambas clases de llamadas (ver sección 4.6). En tal caso la aplicación del modelo M/M/*s* conduce a la predicción de un GoS mejor que el real y a la subdimensión del sistema.

La hipótesis de *cola infinita* será aceptable siempre que el porcentaje de llamadas perdidas por limitación de la capacidad de la cola de espera en el sistema real sea muy bajo, cosa que debe ocurrir si el sistema ha sido dimensionado de forma correcta. Aun cuando se pierdan llamadas por falta de capacidad de la cola, cabe esperar que los usuarios realizarán reintentos hasta conseguir canal, con lo que las llamadas que no entren en la cola física del sistema entrarán en la cola virtual que suponen los reintentos y el resultado será parecido al modelo de cola infinita. Esta situación favorece la hipótesis de cola infinita, ya que a largo plazo todo el tráfico ofrecido al sistema será cursado.

En cuanto a los retardos que pueden producirse en un sistema PMR troncal, la ITU-R considera los siguientes para los sistemas con asignación por mensaje [CCI90]:

- *Retardo de canal*: Es el tiempo de espera en cola hasta obtener canal existiendo posición de despacho libre.

- *Retardo de flota*: Retardo cuando todas las posiciones de despacho de la flota están ocupadas existiendo canal libre.

Por tanto cualquiera de los parámetros del GoS introducidos puede evaluarse en base a uno de los dos criterios anteriores, o considerar la combinación de ambos. Así podemos evaluar la probabilidad de demora del canal (asumiendo posición de despacho libre), la de la flota (asumiendo canal libre) o la probabilidad de que la llamada sea demorada por cualquiera de las dos causas. Es obvio que el retardo de flota no es responsabilidad del operador del sistema por cuanto depende de las posiciones de operadora de que dispone la flota cliente y del tráfico generado por la misma. Por ello la ITU-R centra sus recomendaciones sobre el primero de los dos retardos.

Los parámetros relevantes del GoS calculados de acuerdo al modelo M/M/*s* de la figura 1.3 son la probabilidad de demora *PD* y el tiempo medio de espera en cola o

retardo medio de acceso W(M/M/s), que de acuerdo a la fórmula de Erlang-C [KLE75] pueden ser calculados como:

$$PD = Erlang - C(A,s) = \frac{A^s/s}{\left(1-A/s\right)\left(\sum_{n=0}^{s}\frac{A^n}{n!}+\frac{A^s}{s!}\frac{A/s}{\left(1-A/s\right)}\right)} \tag{1.1}$$

$$W(\mathrm{M}/\mathrm{M}/s) = PD\frac{1/\mu}{s-A} \tag{1.2}$$

donde A representa el tráfico ofrecido al conjunto de los s canales y $1/\mu$ representa la media del tiempo de servicio. Estos valores son independientes de la disciplina de servicio que se siga en la cola. El número medio de unidades en cola está relacionado con el tiempo medio espera a través de la *fórmula de Little* [KLE75]. Este último es un valor muy utilizado en el ámbito de la investigación operativa pero con escaso significado como parámetro del GoS en el ámbito de las telecomunicaciones móviles.

La ITU-R sugiere los objetivos de GoS en base a *percentiles de llamadas* que no superen determinado retardo [CCI90] y dichos percentiles se pueden calcular analíticamente y de forma exacta y sencilla para las disciplinas más habituales de colas, para las cuales existe una solución de la distribución del tiempo de espera en cola: FCFS ("First-Come-First-Served"), Aleatoria ("Random") y LCFS ("Last-Come-First-Served"). Para el modelo de cola M/M/s con disciplina FCFS la probabilidad de que el retardo de acceso supere un tiempo t determinado [TIJ86]:

$$\Pr(W > t) = 1 - W(t) = PD \times \exp\left(-\mu t(s-A)\right) \tag{1.3}$$

Para las disciplinas Random y LCFS (está última es de escaso interés en este entorno de sistemas PMR) puede consultarse [RIO62]. La disciplina FCFS es la más eficiente desde el punto de vista de los percentiles de llamadas demoradas, ya que minimiza los porcentajes de llamadas que sufren una demora mayor que cierta cota, y es por tanto la más utilizada en sistemas PMR [GIB85].

La probabilidad de demora calculada para el modelo M/M/s es una excelente aproximación para la probabilidad de demora en una cola M/G/s [TIJ86] de modo que, aunque la distribución de la duración del mensaje no sea exponencial tal y como se ha comentado en la discusión de las hipótesis, la probabilidad de demora real no variará sensiblemente respecto de la calculada. Cabe citar que la probabilidad de

demora depende en gran medida de la distribución del tiempo entre llegadas, de modo que a menor dispersión del mismo, menor es la probabilidad de demora [SEE85a]. Por otro lado el retardo medio de acceso depende en gran medida de la distribución del tiempo de servicio, y puede variar ostensiblemente si la duración del mensaje no está distribuida conforme a una exponencial negativa. La misma consideración es valida para los percentiles de llamadas demoradas, que dependen de la distribución del tiempo de servicio y de la disciplina que se sigue en la cola (además de la distribución del tiempo entre llegadas en el caso de que ésta no fuera exponencial).

Algunos autores proponen la flota o la posición de despacho en lugar del terminal como unidad básica generadora de tráfico [HAS87, HOA91]. Obsérvese que una flota puede poseer varias posiciones de despacho y por tanto ambas propuestas no tienen por que coincidir necesariamente. Este punto de vista es razonable pensando que en los sistemas PMR la mayoría de las llamadas son a y desde posiciones de despacho. Al ser el número de flotas o de posiciones de despacho muy inferior al número de terminales puede no cumplirse la hipótesis de población infinita que necesita que se cumpla $M>>s$, siendo M el número de fuentes de tráfico. Esta propuesta conduce a un *modelo con población finita* similar al de la figura 1.3 excepto en que la población generadora de tráfico está formada por M fuentes. Si se puede asumir que todas las M fuentes (sean flotas o terminales) generan el mismo tráfico individual, se dice que el sistema es *balanceado* y la solución es la de la cola M/M/*s*//*M*. La ecuación de la probabilidad de demora es conocida como fórmula de Engset-C [HAS87] y conduce a la obtención del retardo medio de acceso.

Si debido a sus características el tráfico que generan las distintas flotas no puede asumirse balanceado el modelo de sistema resultante ya no tiene tratamiento analítico y debe acudirse a la simulación [HAS87] como única herramienta para su evaluación. En [CCI90] se comparan resultados de la eficiencia del canal obtenidos para población finita no balanceada, poblaciones finita balanceada y población infinita.

Algunos autores han estudiado las repercusiones sobre el dimensionado del sistema del hecho de que las llamadas entre móviles utilicen más recursos que las llamadas de despacho en el caso de sistemas no celulares [BAK81] y en el caso de sistemas celulares con móviles en distintas ubicaciones [BAK82], obteniendo fórmulas analíticas derivadas del modelo de Erlang-C para dichas situaciones.

1.5 Sistemas PMR con asignación por mensaje y prioridad

Tal como se ha citado es habitual que los sistemas PMR dispongan de un sistema de prioridades que permita que ciertas llamadas sufran un retardo muy pequeño a costa de un ligero aumento en el retardo de la mayoría. La prioridad debe de estar en manos de pocos usuarios para que sea eficaz: un caso extremo absurdo se produce cuando todos los usuarios son prioritarios y el sistema funciona como si no existieran prioridades. Existen muy diversos métodos de organizar un sistema de prioridades, de los cuales los más sencillos y habituales en sistemas PMR se discuten a continuación.

Hasta la actualidad se han propuesto muchos sistemas diferentes de prioridad que se pueden clasificar de forma genérica en exógenos y endógenos [JAI68]. En los modelos *exógenos* el orden de prioridad no depende del estado del sistema mientras que en los *endógenos* la decisión depende de variables relacionadas con el presente y pasado del sistema. No conocemos casos de prioridad endógena aplicada a telefonía móvil, la razón de ello es probablemente la mayor complejidad tecnológica que se requeriría en el sistema y especialmente en la señalización, no compensada por un mejor funcionamiento.

En los sistemas de *prioridad con robo* o *expulsión* si una llamada intenta acceder al sistema y no encuentra canal disponible, interrumpe (expulsa) una conversación en curso para utilizar el canal que ésta libera al ser interrumpida. Esta forma de prioridad garantiza el acceso inmediato de la llamada prioritaria y degrada la calidad del sistema en la medida en que interrumpe conversaciones en curso, razón por la cual debe reservarse este tipo de prioridad para casos de extrema emergencia. La interrupción de la conversación en un sistema de telefonía es probablemente la degradación de calidad peor tolerada por el usuario. A su vez esta forma de prioridad puede clasificarse y modelarse según que la llamada interrumpida vuelva a ser servida o no; si es servida, debe de tenerse en cuenta si es retomada en el instante donde se interrumpió o por contra se inicia otra vez el servicio. En [BAR95] se encuentra un estudio de los parámetros de calidad de un sistema PMR con prioridad con robo en colas M/M/*s* sin considerar la repetición del servicio de la llamada interrumpida.

Otra forma habitual de prioridad es la *HOL ("Head Of the Line")* mediante la cual una llamada prioritaria que no encuentra canal disponible es ubicada en la cola por delante de todas aquellas con categoría inferior. El acceso inmediato de la llamada prioritaria no está garantizado, pero por otro lado no se interrumpe ninguna llamada

en curso y el retardo de acceso puede llegar a ser muy pequeño. Para la cola M/M/*s* con prioridad HOL y un número determinado de niveles de prioridad existe una solución analítica exacta que se detalla en el apartado 3.2 [GRO74]. Sin embargo para colas en las que la duración de la conversación no está distribuida exponencialmente (M/G/*s*) ya no son válidos dichos resultados, aunque *PD* calculada mediante la ecuación (1.1) sigue siendo una excelente aproximación para la probabilidad de demora: notar que de hecho la probabilidad de demora no se ve afectada por el hecho de que exista prioridad HOL, ya que ésta solamente altera el orden en el que las llamadas son servidas. En el capítulo 3 se presentan aproximaciones para la cola M/G/*s* con prioridad HOL basadas en las ecuaciones (1.1) y (1.2) para los casos en que la distribución del tiempo de servicio es determinística (cola M/D/*s*) [BAR96a] o hipoexponencial [BAR97a]. Este método de prioridad HOL es propuesto en [TEK92] para favorecer las llamadas procedentes de traspasos en sistemas de telefonía móvil pública celular.

Existen otros métodos de prioridad entre los que podemos citar:

- *Prioridad dependiente del tiempo de servicio*: Priorizar las llamadas más breves repercute en un mejor comportamiento del sistema pero el inconveniente es obvio: el sistema desconoce a priori la duración que van atener las llamadas que están esperando en cola. Un modo de solventar este inconveniente es la asignación *Round-Robin* en la que el servidor sirve cada unidad de la cola durante un periodo breve de tiempo especificado. De este modo los servicios de duración menor son priorizados (terminan antes). Esta disciplina tiene sentido en procesos relacionados con ordenadores pero no es aplicable a sistemas de telefonía.

- *Canales reservados (Cutoff Priority)*: Se reservan una serie de canales que únicamente pueden ser utilizados por las llamadas prioritarias. Es un método de prioridad poco eficiente en el cual se produce una infrautilización de los canales dedicados exclusivamente a llamadas prioritarias. Este método ha sido propuesto como forma de favorecer las llamadas procedentes de traspaso ("handover" o "handoff") en sistemas de telefonía móvil pública celular [HON86, TEK91], aunque no tenemos constancia de su utilización práctica. Otros autores han propuesto sistemas mixtos de telefonía móvil de grupo cerrado y pública con diversos esquemas para priorizar una de los clases de llamadas, basados en sistemas de canales reservados [DOG86, POW92]. Actualmente razones legales dificultan la existencia de sistemas privados en los que la conexión a la RTP pueda realizarse de forma automática.

- *Prioridad basada en medidas*: Las llamadas que tienen mayor necesidad de acceder al sistema son favorecidas. Dicha necesidad debe medirse de algún modo, así en [TEK92b] se propone la aplicación de este método a las llamadas procedentes de traspaso en sistemas de telefonía móvil pública (ver sección 1.8).

1.6 Sistemas PMR con asignación por transmisión

En sistemas PMR troncales con gestión por transmisión el canal es asignado únicamente mientras el usuario mantiene el botón PTT ("Push To Talk") presionado, es decir solamente cuando se está transmitiendo [HER93]. Al realizar la asignación de un sistema PMR a nivel de transmisión parece que debe producirse un ahorro espectral al disminuir el tráfico ofrecido al conjunto de canales en alrededor de un 50% (es difícil predecir la reducción exacta debido a la carga que supone la señalización). Sin embargo dicho ahorro no es tan grande como cabría esperar debido a que las exigencias de GoS son mucho más restrictivas cuando la asignación se realiza a nivel de transmisión. De hecho la transmisión acepta una demora muy inferior a la que puede aceptar un mensaje en la obtención del canal. También resulta claro que en este caso por ser la gestión más compleja, la cantidad de señalización que necesita el sistema es mayor, hecho que contribuye también a disminuir la eficiencia.

Un sistema PMR con asignación por transmisión que utilizara el modelo de tráfico de la figura 1.3 de cola M/M/s resultaría tremendamente incómodo para los usuarios que deberían esperar a recibir algún tipo de realimentación indicando cuando pueden empezar a transmitir cada vez que una solicitud de transmisión abandona la cola para pasar a ser servida. El sistema TETRA opta por recortar la transmisión durante el tiempo de espera para la obtención de canal [ETR96], con lo cual no es necesario que el usuario esté pendiente de cuando su intento de transmisión puede ser cursado. Sin embargo el recorte de la parte inicial de la transmisión debido a la posible demora en el acceso será percibido como una importante degradación de la calidad. Por ello los objetivos del GoS deben ser muy restrictivos siendo muy bajos tanto la probabilidad de demora como el retardo medio de acceso (este último es igual al recorte medio que sufren las transmisiones en el acceso a la red). En un sistema con asignación por mensaje el usuario puede tolerar varios segundos de espera en el acceso del mensaje, cantidad que resulta excesiva en el acceso de la transmisión. Por ello la mejora de eficiencia no es tan grande ni clara como cabría esperar.

El modelo de tráfico a aplicar es en este caso el de cola M/M/∞ o modelo de Poisson [KLE75, DES85] (también conocido como de "Blocked-Calls-Held" o de

Molina) siendo el parámetro de calidad de interés (GoS) el recorte producido en las transmisiones. Para probabilidades de demora bajas que son las que deben ocurrir si el sistema está correctamente dimensionado, los cálculos realizados para la cola M/M/*s* coinciden prácticamente con el modelo de Poisson [RIO62, HIL79]. La principal objeción que se puede realizar a la utilización de este modelo es el hecho de que la duración del servicio, en este caso la duración de la transmisión, no está distribuida exponencialmente en un sistema real. Medidas presentadas en el capítulo 4 muestran como la duración de la transmisión vocal en sistemas PAMR está distribuida de forma hiperexponencial de modo que la aplicación del modelo de Poisson tiende a infravalorar los recortes iniciales que se producen en las transmisiones, o lo que es lo mismo a subdimensionar el sistema.

Algunos sistemas como el standard TETRA contemplan también un modo de asignación mixta [HER93, ETR96] que consiste básicamente en una asignación por transmisión, en la que al finalizar el servicio o transmisión el canal se mantiene asignado durante un breve periodo de tiempo adicional ("hangover"). De este modo si el silencio es breve el canal se mantiene asignado para la transmisión siguiente. De este modo se ahorra señalización a costa de disminuir la eficiencia de los canales y por otra parte se minimiza la probabilidad de que una conversación sea cortada por ausencia de canal al iniciarse una de las transmisiones que la componen.

La existencia de un sistema de prioridades a nivel de transmisión no tiene el sentido claro y determinante que tiene en el caso de prioridades de llamadas. La prioridad en asignación por transmisión mejoraría la calidad auditiva percibida por el oyente, pero no el retardo de acceso.

1.7 Sistemas PMR con voz y datos

La tecnología actual de los sistemas PMR basados en el standard MTP1327, y más la tecnología de segunda generación TETRA, favorecen el crecimiento de los servicios de datos. De ellos el servicio de mensajes cortos tiene un interés especial, ya que es conocido el gran mercado existente y que actualmente es suministrado por servicios de radiobúsqueda ("pagers"). Este servicio puede incluso en algunos casos presentarse como sustituto del servicio de voz a un coste muy inferior dado que los requisitos de tiempo son más relajados al tratarse de datos (para casos en que solamente sea necesario dejar un recado como una dirección para flotas de taxis, etc.).

En un la evaluación de un sistema PMR que integre voz y datos y que realice asignación por mensaje deberá de tenerse en cuenta que la distribución del tiempo de

ocupación del canal no será exponencial, ni tampoco hipoexponencial (en el capítulo 2 se caracteriza tráfico solamente de voz). Al mezclarse dos tipos de mensaje con tiempos medios muy diferentes (mucho más cortas las llamadas de datos que las de voz) y en proporción cualquiera, la duración de ocupación del canal estará distribuida de forma hiperexponencial. Esta situación queda descrita en la sección 4.6 con más detalle.

Si el sistema realiza la asignación por transmisión, la voz debería de tener prioridad sobre los datos. De hecho el retardo infringido a la voz repercute en un recorte de la parte inicial de la transmisión, con la consecuente degradación de la calidad percibida. Por otra parte los datos pueden tolerar cierto retardo. En ambos casos los parámetros de calidad de interés son diferentes para voz y para datos. Así para el tráfico de voz el parámetro de interés más importante es el retardo de acceso en asignación por mensaje, y el recorte en asignación por transmisión. Para el tráfico de datos el parámetro de interés es siempre el retardo medio una vez asumida una cierta probabilidad de error suficientemente baja.

1.8 Telefonía móvil pública

Tradicionalmente se han modelado los sistemas de telefonía pública fija como sistemas M/M/*s*/*s*, es decir sistemas de pérdidas ("Blocked-Call-Cleared")que se resuelven mediante la ecuación de Erlang-B. Este modelo no tiene en cuenta los reintentos que se producirán por parte de las llamadas que han encontrado el sistema bloqueado. Dichos reintentos ocurrirán porque el usuario no dispone de una red alternativa a la que pueda optar si encuentra bloqueo en el primer intento. Los reintentos provocan en definitiva que el tráfico cursado sea prácticamente igual al ofrecido, razón por la cual en algunos entornos se utilizan los modelos M/M/*s* (Erlang-C) o M/M/∞ ("Poisson") para dimensionar los sistemas de telefonía fija [SHA90]. Existe también un modelo de Erlang-B con reintentos cuyos resultados son muy parecidos al modelo de Poisson [SHA90]. En una red de telefonía móvil pública disminuye el número de reintentos debido a la opción alternativa que representa la telefonía fija, especialmente las cabinas telefónicas, pero dicha disminución será en cualquier caso pequeña.

La primera diferencia que se observa entre un sistema de telefonía móvil celular y uno de telefonía fija es que la duración de ocupación del canal físico ya no coincide con la duración de la llamada. Los traspasos hacen que en general el canal esté ocupado solamente durante una fracción de la duración de la llamada, pasando luego la misma llamada a otro canal en otra célula. Esta razón hace difícil justificar

una duración de ocupación de canal exponencial, aunque son muchos los autores que la utilizan dadas las facilidades que ello supone en los cálculos analíticos [LEE86b, HON86, GUE87, STE92]. En el capítulo 4 de esta tesis se presentan medidas que muestran que la ocupación de canal está distribuida de forma hiperexponencial, con lo que los diseños realizados en base a una distribución exponencial en general resultarán subdimensionados. En este apartado mantendremos la distribución del tiempo de servicio más comúnmente aceptada en la literatura, es decir, la exponencial, a pesar de las carencias que hemos puesto de manifiesto.

En el modelo de un sistema de telefonía móvil pública debemos considerar las llamadas nuevas y las procedentes de traspaso. Las llamadas procedentes de traspaso generan reintentos con un periodo muy breve comparado con la duración de ocupación del canal y con el periodo de reintento de las llamadas nuevas, debido a que la petición de traspaso se realiza de forma automática por necesidades de calidad o tráfico. De este modo podemos modelar los traspasos ofreciéndose a una cola de capacidad infinita, y las llamadas nuevas ofreciéndose a los mismos servidores pero sin cola, es decir en modo de pérdidas. Este es un modelo sencillo tomado de [TEK92b] y que se representa en la figura 1.4; el proceso de Markov asociado al modelo se representa a su vez en la figura 1.5. La cola de la figura 1.4 representa una cola virtual que no existe físicamente y cuya disciplina de servicio será en general aleatoria. Los tráficos A_N y A_H representan los tráficos ofrecidos de nuevas llamadas y de traspasos ("handover") respectivamente. A_T representa el tráfico total ofrecido al sistema, suma de los tráficos citados. La misma notación es utilizada para las tasas de llegadas al sistema en la figura 1.5.

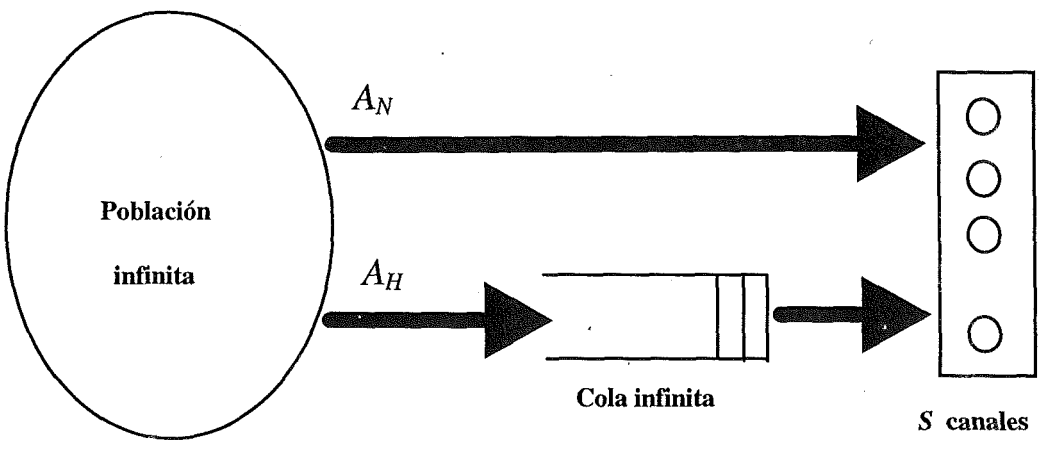

Figura 1.4. Modelo de un sistema de telefonía móvil pública incluyendo llamadas nuevas y procedentes de traspaso.

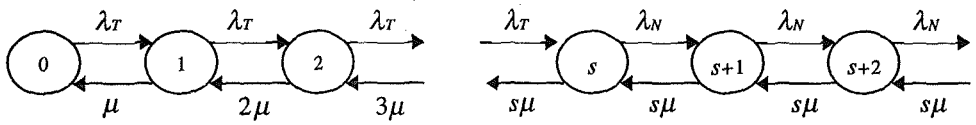

Figura 1.5 Diagrama de transición de estados del modelo de la figura 1.4.

La solución del proceso de Markov de la figura 1.5 conduce a las siguientes probabilidades para cada uno de los estados (n representa el número de unidades en el sistema, tanto en la cola como en los canales):

$$\begin{aligned} P_n &= \frac{A_T{}^n}{n!} P_0 \qquad \text{para} \quad n \le s \\ P_n &= \left(\frac{A_H}{s}\right)^{n-s} P_s \qquad \text{para} \quad n > s \end{aligned} \tag{1.4}$$

$$P_0 = \frac{1}{\left(\sum_{n=0}^{s} \frac{A_T{}^n}{n!} + \frac{A_T{}^s}{s!} \frac{A_H/s}{\left(1 - A_H/s\right)}\right)} \tag{1.5}$$

De las probabilidades de estado puede obtenerse la probabilidad de bloqueo PB, que es la probabilidad de que una llamada encuentre todos los canales ocupados igual a la probabilidad de que un traspaso sea demorado.

$$PB = \sum_{n=s}^{\infty} P_n = P_s \frac{1}{1 - \rho_H} \tag{1.6}$$

Otro parámetro de interés del GoS es el retardo medio de acceso para un traspaso W (las nuevas llamadas no pueden sufrir retardo).

$$W = \frac{\overline{N_q}}{\lambda_H} = \frac{\sum_{n=s+1}^{\infty} (n-s) P_n}{\lambda_H} = PB \frac{1/\mu}{s - A_H} \tag{1.7}$$

Este modelo no recoge la posibilidad de que un traspaso no pueda realizarse debido a que el terminal sale de la zona de cobertura antes de que exista un canal libre. Una forma de incluir este efecto es asumir una cola finita con disciplina FCFS en lugar de la cola infinita de la figura 1.4 [TEK92b]. Los mismos autores han

propuesto modelos con disciplinas de cola dependientes de la necesidad que tiene el terminal móvil de ser traspasado. Dicha necesidad puede ser medida desde las estaciones base en función de la potencia que están recibiendo del terminal móvil: a menor potencia mayor necesidad de que el móvil realice el traspaso a otra celda. Este es el llamado esquema de *priorización basado en medidas* que proporciona mejores resultados que el modelo de la figura 1.4. Los estudios analíticos que incluyen estos efectos son mucho más complejos que el modelo presentado aquí, y dependen de la movilidad de los terminales (velocidad, dirección y distribución geográfica de los móviles dentro de la celda).

La complejidad del modelo utilizado puede aumentar prácticamente sin límite. Así en [RAP93] el autor presenta un modelo en el que el tiempo de ocupación del canal está distribuido conforme a una combinación de exponenciales negativas (en lugar de la exponencial negativa simple) y en el que considera distintas plataformas de movilidad (diferentes grupos de usuarios se mueven conforme a patrones diferentes) además de otras hipótesis habituales en este tipo de estudios (ver sección 4.4). La solución analítica exacta para los parámetros del GoS para un sistema con 15 canales por celda requiere 4 minutos de CPU mientras que con 25 canales requiere de unos 45 minutos de CPU (siempre utilizando un supercomputador).