UNIVERSIDAD POLITECNICA DE VALENCIA

DEPARTAMENTO DE COMUNICACIONES

Máster Universitario en Tecnologías, Sistemas y Redes de Comunicaciones

Análisis Aproximado de Redes de Telecomunicaciones Basado en la Separación de Escalas Temporales

Autor: Luis Tello Oquendo

Director 1: Vicent Pla Boscà

Director 2: Jorge Martínez Bauset

Fecha de comienzo: 25/02/2013

Lugar de trabajo: Grupo de Interconexión de Redes de Banda Ancha - Departamento de Comunicaciones

Objetivos – El objetivo principal de este trabajo lo constituye la evaluación de la recientemente propuesta generalización de la aproximación cuasi-estacionaria introducida en [1] para el análisis del desempeño de redes de telecomunicaciones desde la perspectiva de tráfico. Para ello, se consideran dos sistemas diferentes: un sistema cognitive radio (CRS) con distintos tipos de usuarios y un sistema de servicios integrados (ISS) con distintos tipos de tráfico. Se realiza un estudio comparativo de la precisión y coste computacional al utilizar tanto el método clásico de aproximación denominado aproximación cuasi-estacionaria (QSA) como la aproximación cuasi-estacionaria generalizada (GQSA) propuesta en [1]; en particular se estudia el efecto de variar la separación de escalas temporales de cada sistema en la precisión de los métodos de aproximación. También se evalúa el compromiso entre precisión y coste computacional de los dos métodos.

Metodología – Para la realización de este trabajo, se han modelado tanto el CRS como el ISS utilizando procesos de Markov. Se ha obtenido la solución exacta de los parametros que se utilizan para evaluar el rendimiento de cada sistema. Se han implementado los métodos QSA y GQSA para estimar los parámetros de rendimiento en cada sistema. Mediante la utilización de un parámetro de GQSA denominado radio, y de un factor multiplicativo que lo denominamos factor de aceleración de eventos (f) que permite controlar el grado de la separación de escalas temporales, se ha estudiado la evolución en el comportamiento de la precisión de GQSA. Se ha calculado el error relativo en los parámetros de rendimiento analizados y se han medido los tiempos de ejecución del algoritmo para el cálculo de los parámetros de rendimiento.

Desarrollo de prototipos y trabajo de laboratorio – Como punto de partida se estudió la metodología de separación de escalas temporales. Se trasladó este concepto a la dinámica de los elementos que conforman tanto el sistema de servicios integrados, como el sistema de cognitive radio, dando como resultado modelos de Markov bi-dimensionales para cada sistema. Una vez modelados los sistemas, se realizó el analisis exacto de los mismos; en los que se evidenció claramente, con tamaños de sistemas grandes, el elevado coste computacional requerido para encontrar la solución. Posteriormente, se analizó y replicó los resultados del trabajo realizado en [1] donde se propone la generalización de la aproximación cuasi-estacionaria. Partiendo del trabajo realizado en [2], se procedió a implementar GQSA en el sistema de cognitive radio, para la estimación de los parámetros que se utilizan para evaluar el desempeño del sistema. Una vez implementado GQSA tanto en ISS como en CRS, se realizó un estudio exhaustivo de este método en los dos sistemas, debido a sus diferencias cualitativas y a que se encontró un comportamiento no-monótono de la aproximación al evaluar los parámetros de rendimiento medidos. Se estudió la evolución tanto de GQSA como de QSA en función de la separación de escalas temporales. Se midieron los tiempos de ejecución y se compararon los resultados con los de la solución exacta de los sistemas, con la finalidad de evaluar el compromiso entre precisión de los métodos de aproximación y coste computacional al implementarlos en los sistemas.

Resultados – Como se esperaba, tanto en CRS como en ISS, con cualquier configuración de tamaño y carga, a mayor separación de escalas temporales en los flujos de tráfico, los

valores aproximados de los parámetros de rendimiento evaluados tienden a los valores exactos. A medida que la separación de escalas temporales va disminuyendo, los valores aproximados van separándose de los valores exactos, comenzando con aquellos en los que se utilizó un menor radio para el análisis. Esto implica que para obtener una precisión elevada a diferentes escalas temporales, es necesario analizar el sistema con un radio elevado, incrementando el coste computacional. Se ha observado que la tendencia general de GQSA es acercarse gradualmente al valor exacto con cada incremento de radio utilizado en la aproximación. No obstante, en algunos escenarios se puede lograr una precisión elevada incrementando el radio ligeramente. También se descubrieron ciertos casos especiales en los que incrementar el radio inicialmente empeora la precisión. El hecho de saber cuando el error relativo decrece, y cuando no, al incrementar el radio, depende de un modo complejo de varios factores, lo que hace difícil de predecir en qué casos la precisión se puede mejorar al utilizar GQSA respecto a QSA. Un hallazgo inesperado es que, en ciertos casos específicos, GQSA resulta ser una buena aproximación no sólo para el regimen cuasi-estacionario, sino también para el regímenes considerablemente alejados del cuasi-estacionario.

Líneas futuras – En vista de los resultados y limitaciones del método introducido en [1], se plantea explorar y desarrollar generalizaciones alternativas porque es crucial saber en qué condiciones se mejora la precisión. También es necesario indentificar un test o parámetro con el que se pueda decir a priori (dar una cota o estimación del error) si va a ser útil o no utilizar la aproximación; i.e., establecer un criterio alternativo que dijese la fiabilidad de la aproximación, sin necesidad de realizar el cálculo exacto.

Publicaciones – "Approximate Analysis of Wireless Systems Based on Time-Scale Decomposition", 6ta. edición del Congreso *IFIP Wireless Days 2013*, ACEPTADO.

Abstract – Markov chains are a widely used modeling tool for communication networks. The system size and the existence of different user types often make the analysis of the Markov chain computationally intractable. When the events of each user type occur at sufficiently separated time scales, the so-called quasi-stationary approximation (QSA) has proven to be accurate and highly efficient. Recently, a generalization of the quasi-stationary approximation (GQSA) has been introduced. The new approximation aims to improve the accuracy at the price of higher computational cost. In this paper, we carry out a comparative study of the accuracy and computational cost of both approximation methods QSA and GQSA. In particular, we explore the evolution of accuracy as the separation between time scales varies, and the trade-off between accuracy and computational cost. Our results indicate that while the new GQSA improves the accuracy in some instances, it does not occur in all of them; and more importantly, it is difficult to predict in which cases accuracy can be enhanced by the new method.

Autor: Luis Tello Oquendo, email: luiteloq@teleco.upv.es
Director 1: Vicent Pla Boscà, email: vpla@dcom.upv.es
Director 2: Jorge Martínez Bauset, email: jmartinez@upvnet.upv.es
Fecha de entrega: 09-09-13

Índice

1. Introducción

Debido a la naturaleza compleja del tráfico en las redes de telecomunicaciones, comúnmente se han utlizado cadenas y procesos de Markov como herramientas para modelar sistemas de comunicaciones con la finalidad de estudiar su desempeño. De esta manera, se puede ofrecer un servicio de calidad dentro de las restricciones físicas, tecnológicas o económicas a los que esté sujeto el sistema. Una de las principales ventajas de la utilización de modelos de Markov es que es lo suficientemente general para capturar los factores dominantes de la incertidumbre del sistema en análisis, permitiendo conocer su comportamiento y evolución de forma analítica.

Actualmente, la mayoría de sistemas son dinámicos e inevitablemente grandes y complejos. En muchas ocasiones, los diferentes elementos de un gran sistema evolucionan a diferentes velocidades. Algunos de ellos varían rápidamente y otros cambian lentamente. Un sistema dinámico evoluciona como si los diferentes componentes utilizaran diferentes relojes o escalas de tiempo. Para describirlos cuantitativamente, es crucial decidir el orden de magnitud de las tasas para los diferentes elementos a través de comparaciones. Debemos tener en mente que lo *rápido* vs. *lento* y *mucho tiempo* frente a *corto tiempo* son términos relativos. De hecho, la separación de escalas de tiempo es a menudo inherente a los problemas de fondo.

El hecho de incorporar todos los factores importantes de un sistema en su correspondiente modelo, a menudo resulta en un gran espacio de estados de la cadena o proceso de Markov correspondiente; esto implica que el coste computacional para calcular su desempeño se incrementa en gran medida.

Para reducir la complejidad, se ha sugerido un enfoque jerárquico, lo que conduce a una formulación en dos escalas de tiempo [3]. El enfoque jerárquico se basa en la descomposición de los estados de las cadena o proceso de Markov asociado al sistema en varias clases o, posiblemente, varias clases más un grupo de estados transitorios. La esencia es que dentro de cada clase las interacciones son fuertes y entre las diferentes clases las interacciones son débiles.

Considerando esta descomposición en escalas de tiempo, se han desarrollado técnicas de aproximación para reducir el coste computacional. Una de estas aproximaciones computacionalmente eficiente, basada en la separación de escalas temporales y que se

utiliza a menudo, es la aproximación cuasi-estacionaria (QSA) [4, 5, 6].

En [1] los autores introducen un nuevo método, también basado en la descomposición de escalas temporales, llamado aproximación cuasi-estacionaria generalizada (GQSA), que proporciona una manera de equilibrar la complejidad computacional y la precisión de la aproximación. Los autores aplican este nuevo método en un sistema de servicios integrados (ISS) que sirve tráfico streaming o real-time (RT) y tráfico elástico o non-real time (NRT), sin considerar el efecto que produce variar la separación de escalas temporales en la precisión de la aproximación.

Utilizando éste método, se evalúa su comportamiento en un sistema cognitive radio (CRS) que, a nivel de modelo, presenta importantes diferencias cualitativas con respecto a los recursos disponibles para cada tipo de usuario y las tasas de servicio en cada estado del modelo como se describe en la sección 2. El objetivo es explorar GQSA en ambos sistemas: CRS and ISS.

Por la importancia que tiene diseñar sistemas los más efectivos posible en relación a sus costes, con un grado de servicio predefinido, para lo cual hay que conocer la demanda futura de tráfico y la capacidad de los elementos del sistema, se aborda el problema desde el punto de vista de tráfico. Se desarrollan dos modelos analíticos para evaluar el desempeño de estos sistemas: uno para CRS y otro para ISS. En estos modelos se considera que la dinámica de los tipos de usuarios (en el CRS) o tipos de tráfico (en el ISS) operan en escalas de tiempo suficientemente separadas, lo que permite utilizar métodos de aproximación basados en la descomposición de escalas temporales para simplificar los cálculos.

En cuanto a CRS, el concepto de Cognitive Radio propone impulsar la utilización del espectro al permitir que los usuarios cognitivos (usuarios secundarios, SU) acceder a canales inalámbricos con licencia de manera oportunista por lo que la interferencia para los usuarios con licencia (usuarios primarios, PU) se mantiene al mínimo [7]. Dado que en un CRS hay diferentes tipos de usuarios, el tamaño del espacio de estados se incrementa rápidamente con el número de canales que estos requiren, lo que implica el crecimiento de la complejidad computacional relacionada con la solución del modelo de Markov asociado con el sistema. Por esta razón, típicamente se requiere utilizar una aproximación.

Por otro lado, en un ISS, se espera que futuras generaciones de redes de banda ancha soporten una gran variedad de aplicaciones, por lo general agrupadas en dos grandes cate-

gorías: en real-time (RT) (por ejemplo, voz y vídeo) y non-real time (NRT) (por ejemplo, navegación web, correo electrónico y transferencia de archivos) [8]. Cuando el número de canales tanto para tráfico RT como para flujos NRT es grande, la complejidad de los cálculos relacionados con la solución del modelo de Markov asociado con el sistema se vuelve prohibitivo. Por lo tanto, se requieren aproximaciones computacionalmente eficientes [1].

La contribución del presente trabajo, se puede resumir de la siguiente manera: en primer lugar, se evalúa GQSA aplicándolo a un sistema diferente al utilizado en [1]: CRS. En segundo lugar, en ambos sistemas (CRS y ISS) se evalúa el comportamiento de la aproximación cuando varía la separación de escalas temporales desde el régimen de cuasi-estacionario hacia el régimen fluido. En tercer lugar, se analiza el compromiso entre la precisión y el coste computacional de los métodos de aproximación basados en la separación de escala temporales.

El documento sigue la siguiente estructura: en la sección 2 se describen las características de los sistemas (CRS y ISS) y se explica la generación de los modelos de Markov para cada uno de los casos. Esto incluye los parámetros de tráfico que definen el comportamiento de los usuarios, el sistema y las prestaciones. La sección 3 presenta las aproximaciones QSA y GQSA basadas en la separación de escalas temporales para simplificar los cálculos y la complejidad computacional que se deriva al realizarlos. En la sección 4 se detalla la evaluación numérica de las aproximaciones y se muestran los resultados de los parámetros medidos que han sido utilizados para evaluar el desempeño de los sistemas. También se evalúa la precisión y el coste computacional de las aproximaciones. Finalmente, la sección 5 recoge las conclusiones del trabajo.

2. Descripción de modelos y análisis exacto de los sistemas

El alcance y propósito del desempeño de un sistema hoy en día pueden ser analizados antes de llevarlo a la práctica en base a la evaluación de la información disponible. Existen dos estrategias que pueden ser tomadas en consideración con el propósito de conseguir la evaluación de las prestaciones de un sistema dado. La primera estrategia es la simulación y la otra es la solución analítica. La solución analítica se basa usualmente en el uso de cadenas y procesos de Markov de los cuales, las probabilidades de estado estacionario, pueden ser determinadas con base a mediciones de rendimiento. En esta sección, se detallan las características de los sistemas a analizar y se describen los modelos de Markov asociados con los mismos.

Una de las características más destacadas de este trabajo es el uso de múltiples escalas de tiempo en Procesos de Markov. Intuitivamente, no todos los elementos o componentes de un sistema a gran escala evolucionan al mismo ritmo. Algunos de ellos cambian rápidamente y otros varían lentamente. Los diferentes tipos de variaciones nos permiten reducir la complejidad a través de la descomposición y la agregación.

Una manera eficaz de tratar la descomposición es la separación de escalas de tiempo, por lo que se ha configurado los sistemas como si hubiera dos escalas de tiempo: rápido versus lento. Con la finalidad de llevar a cabo la separación de escalas temporales, se utilizan modelos de Markov bi-dimensionales (considerando las dos escalas de tiempo).

De esta manera, para el CRS en régimen cuasi-estacionario, la ocurrencia de eventos –llegadas o salidas al sistema– de los SUs es rápida en comparación a la de los PUs; de igual forma para los tipos de tráfico que conforman el ISS: en régimen cuasi-estacionario, la ocurrencia de eventos del tráfico NRT es rápida en comparación a la del tráfico RT.

2.1. Sistema Cognitive Radio

De la misma manera que en [2], se modela el tráfico de PU y SU al nivel de sesión (conexión) y se ignora interacciones a nivel de paquetes (scheduling, buffer management, etc.). Se asume una capa MAC ideal para SU_s, lo que permite la compartición perfecta de los canales asignados entre los SU_s activos (todos las SU_s activos pueden conseguir la

misma porción de ancho de banda), introduce retardo nulo y sus mecanismos de control consumen cero recursos. Adicionalmente, también se supone que un *SU* activo puede sensar la llegada de un *PU* en el mismo canal instantáneamente y fiablemente. De esta manera, los parámetros de rendimiento obtenidos pueden ser considerados como una cota superior.

El sistema cognitive radio tiene C_1 canales primarios (PC_s) que pueden ser compartidos entre usuarios primarios y secundarios, y C_2 canales secundarios (SC_s), disponibles solo para usuarios secundarios (sin licencia); $C = C_1 + C_2$ es el número total de canales en el sistema. Se debe tener en cuenta que los SC_s se pueden obtener de, por ejemplo, bandas sin licencia, como se propone en [9]. Este supuesto es aplicable a un escenario de implementación de *coexistencia* para CRNs. Por otra parte, ya que podría ser de interés comercial para las redes primarias y secundarias a *cooperar*, los canales secundarios se pueden obtener basándose en un acuerdo con la red primaria [10].

Un *SU* en los PC_s podría verse obligado a abandonar su canal si un *PU* confirma que pretende iniciar una nueva sesión. Como los SU_s soportan *spectrum handover*, un *SU* desalojado puede continuar con su comunicación que cursa si existe un canal libre disponible. De otra manera, es *forzado a terminar*. A la llegada de un *SU*, éste selecciona un canal libre disponible con igual probabilidad entre PC_s y SC_s. Con el uso de un esquema de selección que elige un SC libre como primera opción, y recurre a ocupar un PC sólo cuando todos los SC_s están ocupados, se reducirá la interferencia causada a los PU y la tasa de spectrum handovers.

Con motivo de tratabilidad matemática, se asumen llegadas de Poisson y tiempos de servicio distribuidos exponencialmente tanto para PU_s como SU_s. Las tasas de llegada para sesiones *PU* y *SU* son λ_1 y λ_2 respectivamente; las tasas de servicio son μ_1 y μ_2 y requieren el consumo de un canal cuando son aceptados en el CRS, con lo que se garantiza que no hay solapamiento entre ningún usuario. Se denota por (i, j) el estado del sistema, cuando hay i sesiones en curso de *PU* y j sesiones en curso de *SU*. El conjunto de posibles estados en el sistema es

$$\mathcal{S} := \{(i, j) : 0 \leq i \leq C_1, 0 \leq j \leq C_2, 0 \leq i + j \leq C\}$$

y la cardinalidad de $\mathcal{S}$ se define como

$$|\mathcal{S}| = (\frac{C_1}{2} + C_2 + 1) \cdot (C_1 + 1).$$

El diagrama de transición de estados del sistema se representa en la Figura 1.

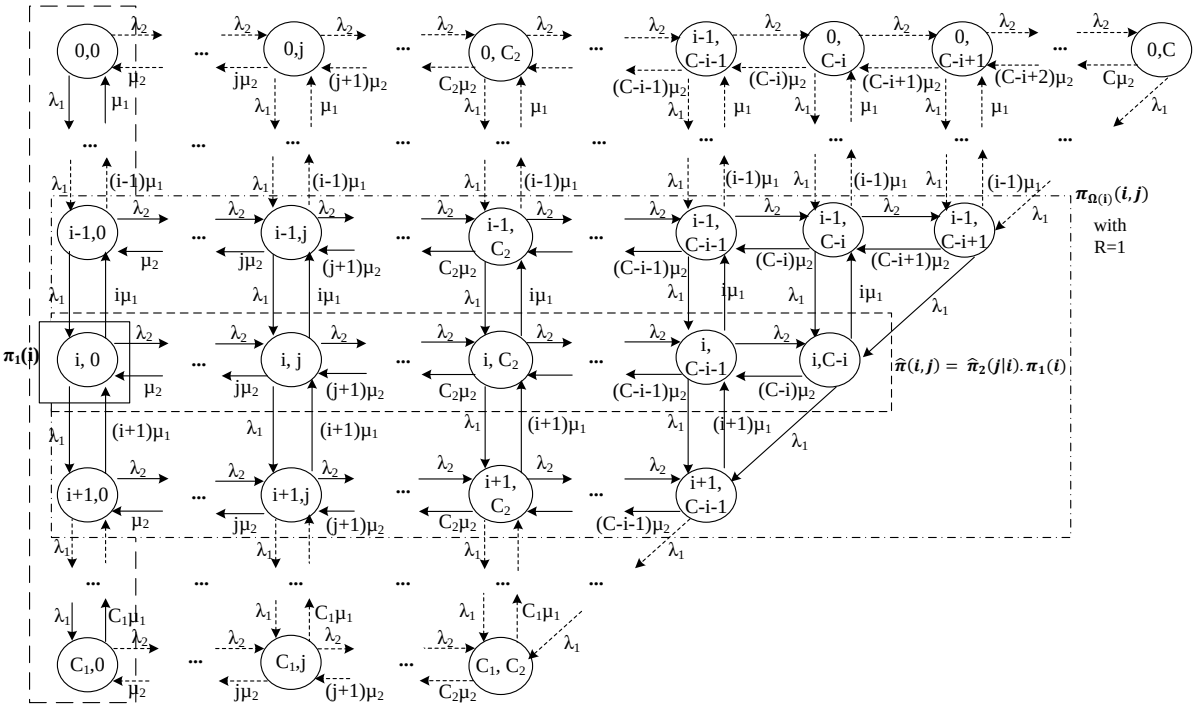

Figura 1: Diagrama de transición de estados, Sistema Cognitive Radio.

Dado el conjunto de posibles estados y sus trancisiones en el modelo de Markov, se pueden construir las ecuaciones de balance global y la ecuación de normalización. Con estas ecuaciones se procede a calcular las probalilidades de estado estacionario, que se denotan como $\pi(i, j)$.

Para determinar los parámetros que sirven para evaluar el rendimiento del sistema, se utilizan las siguientes ecuaciones:

$$P_1 = \sum_{k=0}^{C_2} \pi(C_1, k) \quad , \quad P_2 = \sum_{k=C_2}^{C} \pi(C - k, k), \tag{1}$$

$$P_{ft} = \frac{\lambda_1(P_2 - \pi(C_1, C_2))}{\lambda_2(1 - P_2)}, \tag{2}$$

$$Th_2 = \sum_{j=1}^{C} \sum_{i=0}^{\alpha} j\mu_2 \cdot \pi(i, j), \tag{3}$$

donde P_1 es la probabilidad de bloqueo de PU_s, que claramente coincide con la obtenida utilizando un modelo de pérdidas Erlang-B con C_1 servidores; P_2 es la probabilidd de bloqueo se SU_s, i.e. la fracción de sesiones de SU rechazadas a su llegada debido a que encontraban el sistema completo; P_{ft} es la probabilidad de terminación forzada de SU_s, i.e. la tasa de sesiones de SU forzadas a terminar dividida por la tasa de sesiones de SU aceptadas; Th_2 es el throughput de SU_s, i.e. la tasa de sesiones de SU que han sido completadas satisfactoriamente y por último en (3) se define $\alpha = \text{mín}(C_1, C - j)$.

2.2. Sistema de Servicios Integrados

Se utiliza el mismo modelo definido en [1] para un sistema de servicios integrados que sirve tráfico real-time (RT) y tráfico non-real time (NRT). Se considera un enlace cuyos recursos limitados (C Mbps en total) son compartidos entre peticiones RT y NRT. Al tráfico RT se le da estricta prioridad sobre el tráfico NRT. Inicialmente se asume que todas las llamadas RT son de la misma clase y cada una requiere un canal de tasa c b/s durante toda la duración del servicio para satisfacer la calidad de servicio requerida.

Se denota por N_{rt} el número máximo de canales para llamadas RT. Cuando llega una llamada RT, ésta ocupa un canal, si hay disponible, caso contrario la llamada es bloqueada. Se ha establecido N_{rt}, de tal manera que $N_{rt} \cdot c$ es suficientemente menor que C para evitar la inanición del tráfico NRT.

Sea $n_{rt}(t)$ el número de llamadas RT en un tiempo t, $t \geq 0$; entonces $\{n_{rt}(t), t \geq 0\}$ es el proceso de RT. Los flujos NRT son servidos uniformemente por la capacidad sobrante del tráfico RT de acuerdo con una disciplina processor sharing (PS). Sea $n_{nrt}(t)$ el número de flujos NRT en el sistema en un tiempo t, $t \geq 0$; entonces, $\{(n_{rt}(t), n_{nrt}(t)), t \geq 0\}$ es el proceso RT y NRT en conjunto. La capacidad disponible para todo el tráfico NRT en un tiempo t está dado por $C_{nrt}(t) = C - n_{rt}(t) \cdot c$.

El bit-rate de cada flujo NRT admitido en un tiempo t viene dado por $c_{nrt}(t) = C_{nrt}(t)/n_{nrt}(t)$, el cual se actualiza con cada llegada o salida de llamadas RT o flujos NRT admitidos. Para satisfacer la calidad de servicio de los flujos NRT admitidos, el número máximo de flujos NRT concurrentes se limita a N_{nrt} . Por consiguiente, un flujo NRT que llega en un tiempo t es bloqueado si $n_{nrt}(t) = N_{nrt}$.

Para las peticiones tanto de RT como de NRT se asumen llegadas de Poisson con tasas λ_{rt} y λ_{nrt} respectivamente. El tiempo de servicio de cada petición RT admitida está distribuido exponencialmente y su tasa de servicio es μ_{rt}. Por otra parte, como las sesiones de datos generan tráfico NRT, su tiempo de permanencia en el sistema dependerá de los recursos disponibles. El tamaño de los flujos generados por las sesiones de datos están exponencialmente distribuidos con media L (bits).

Se denota por (i, j) el estado del sistema, cuando hay i llamadas RT en curso y j flujos NRT. El conjunto de posibles estados en el sistema es

$$\mathcal{S} := \{(i, j) : 0 \leq i \leq N_{rt}, 0 \leq j \leq N_{nrt}, 0 \leq i + j \leq N_{rt} + N_{nrt}\}$$

y la cardinalidad de $\mathcal{S}$ está dada por $|\mathcal{S}| = (N_{rt} + 1)(N_{nrt} + 1)$. El diagrama de transición de estados se representa en la Figura 2.

Dado el conjunto de posibles estados y sus transiciones en el modelo de Markov, se pueden construir las ecuaciones de balance global y la ecuación de normalización. Con estas ecuaciones se procede a calcular las probalilidades de estado estacionario, que se denotan como $\pi(i, j)$.

Se debe considerar que la tasa de servicio de los flujos NRT varían de acuerdo a las n_{rt} llamadas RT en el sistema como sigue:

$$\mu_{nrt}^{(i)} = \frac{C - i \cdot c}{L}. \tag{4}$$

Los parámetros para evaluar el rendimiento del sistema, se pueden calcular con las ecuaciones que se detallan a continuación:

$$P_{nrt} = \sum_{k=0}^{N_{rt}} \pi(k, N_{nrt}), \tag{5}$$

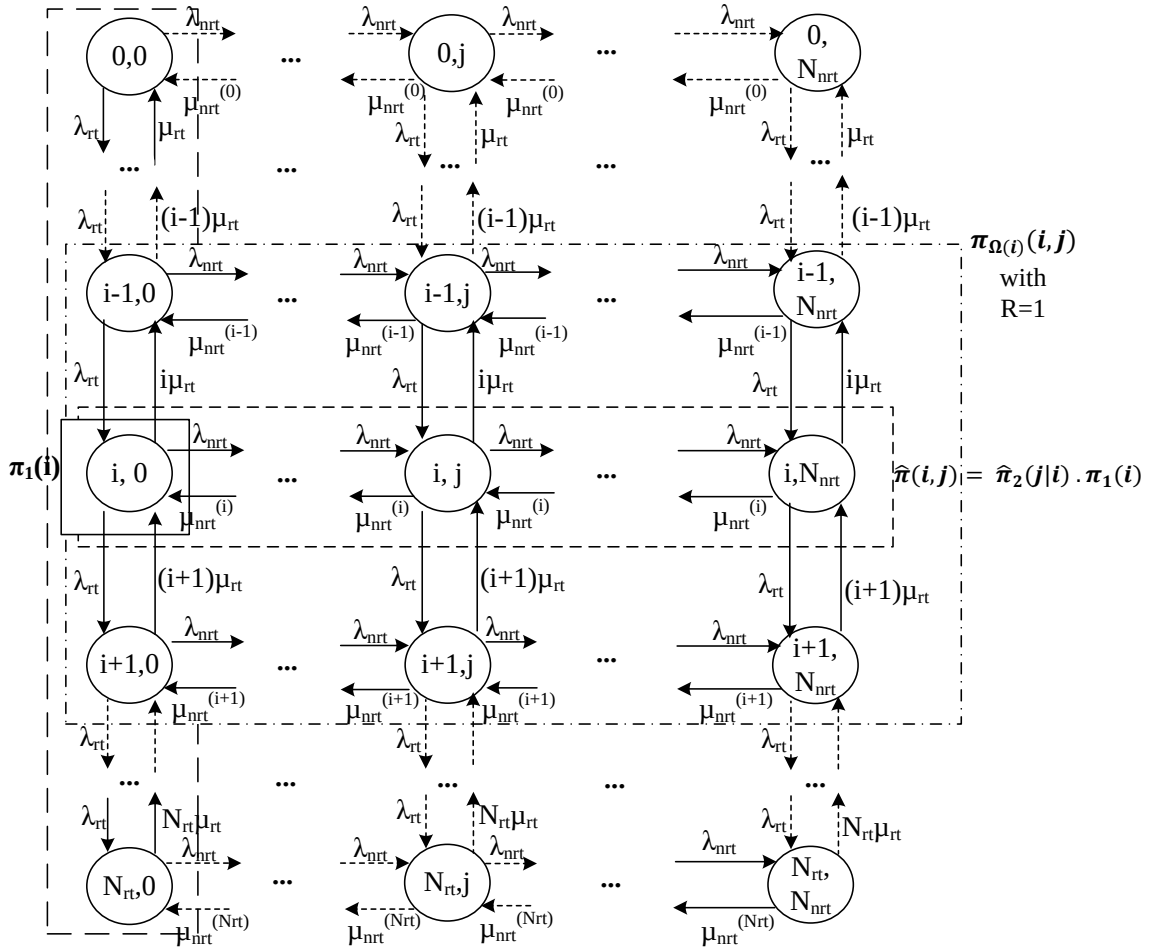

Figura 2: Diagrama de transición de estados, Sistema de Servicios Integrados.

$$E\left[X_{nrt}\right] = \sum_{j=1}^{N_{nrt}} \sum_{i=0}^{N_{rt}} j \cdot \pi(i,j), \tag{6}$$

$$E\left[D_{nrt}\right] = \frac{E\left[X_{nrt}\right]}{\lambda_{nrt}(1 - Pb_{nrt})}, \tag{7}$$

donde P_{nrt} es la probabilidad de bloqueo de un flujo NRT, $E\left[X_{nrt}\right]$ es el número medio de flujos NRT en el sistema y $E\left[D_{nrt}\right]$ es el retardo de transferencia promedio de un flujo NRT.

3. Aproximaciones basadas en la separación de escalas temporales

Desde el punto de vista de modelado, mientras los modelos de Markov han sido desarrollados para el análisis exacto de redes de telecomunicaciones, estos pueden ser numéricamente complicados [4].

Un modelo de Markov con un espacio de estados finito, pero grande, puede hacer el análisis exacto computacionalmente intratable y problemático; por lo que un enfoque de descomposición es a menudo atractivo. Idealmente, nos gustaría dividir el problema subyacente en subproblemas que se pueden resolver de forma totalmente independiente, entonces podemos juntar las soluciones de los subproblemas para obtener la solución a todo el problema.

Considerando una cadena de Markov homogénea; si, por ejemplo, la matriz de transición se puede descomponer en varias sub-matrices de transición (en una forma diagonal por bloques), el problema puede ser resuelto fácilmente utilizando métodos de descomposición [11, 12]. Desafortunadamente, el mundo real no es ideal. En lugar de tener descomponibilidad completa, con frecuencia se encuentra casos casi completamente descomponibles.

Uno de los objetivos principales de este trabajo es el tratamiento de este tipo de modelos a través de la formulación de modelos de Markov en tiempo contínuo con un enfoque de dos escalas de tiempo.

Como la dinámica de los componentes o elementos de los sistemas en análisis (tipos de usuarios –PU o SU– en CRS o tipos de tráfico –RT o NRT– en ISS) operan a escalas de tiempo lo suficientemente separadas, se puede recurrir a aproximaciones altamente eficientes basados en la descomposición de escalas temporales, lo cual permite simplificar grandemente los cálculos [2].

En esta sección, se describe la aproximación cuasi-estacionaria (QSA) y la aproximación cuasi-estacionaria generalizada (GQSA), que reducen la complejidad y el coste computacional para hallar una solución estimada de los parámetros de rendimiento en los sistemas CRS y ISS.

3.1. Aproximación Cuasi-estacionaria –QSA–

La aproximación más simple, que produce resultados fácilmente computables es la llamada aproximación cuasi-estacionaria (o, cuasi-estática) [13].

Tomando como ejemplo el CRS, la aproximación se puede aplicar desacoplando los dos tipos de usuarios del sistema, para analizar sus parámetros de rendimiento independientemente. Se representa cada tipo de usuario del sistema en una dimension del modelo correspondiente. En la Figura 1 se puede ver el proceso de Markov correspondiente a los PUs en el eje y y el proceso de Markov correspondiente a los SUs en el eje x.

Dado que los PUs utilizan los recursos de canal independientemente de la existencia de SUs, el análisis de rendimiento de PU se puede evaluar de forma independiente de una manera exacta. Entonces el rendimiento de SU se puede aproximar.

El proceso de aproximación consiste en dos fases. En la primera fase se obtiene la distribución de probabilidades de estado estacionario para PU, i.e. $\{\pi_1(i) : i = 0, ..., C_1\}$, donde $\pi_1(i)$ es la probabilidad de encontrar i sesiones en curso en un sistema $M/M/C_1/C_1$ únicamente con PUs.

La segunda fase consiste en calcular la distribución de probabilidades de estado estacionario de la sesión SU (j), condicionado al estado (i) de la distribución de PU en la que se encuentre: $\widehat{\pi}_2(j|i)$. También, $\widehat{\pi}_2(j|i)$ es la probabilidad de encontar j sesiones en curso en un sistema $M/M/C-i/C-i$ únicamente con SUs.

Tanto $\pi_1(i)$ como $\widehat{\pi}_2(j|i)$ pueden ser determinados independientemente utilizando recursiones simples, dado que sus correspondientes modelos de Markov son procesos de nacimiento y muerte unidimensionales.

Finalmente, la distribución de probabilidades de estado estacionario de CRS se pueden aproximar mediante:

$$\pi(i,j) \approx \widehat{\pi}(i,j) = \pi_1(i) \cdot \widehat{\pi}_2(j|i). \tag{8}$$

Se puede obtener suficiente precisión con la utilización de QSA. En el régimen cuasi-estacionario (QS), para el CRS, la probabilidad de bloqueo de PUs es $P_1^{qs} = \pi_1(C_1)$; la probabilidad de bloqueo de usuarios secundarios es P_2^{qs}, el throughput de SUs is Th_2^{qs} y la probabilidad de terminación forzada de SUs es P_{ft}^{qs}. Estos parámetros se pueden calcular

empleando la distribución definida en (8) de la siguiente manera [2]:

$$P_2^{qs} = \sum_{k=c2}^{C} \widehat{\pi}(C-k,k), \tag{9}$$

$$P_{ft}^{qs} = \frac{\lambda_1(P_2 - \widehat{\pi}(C_1,C_2))}{\lambda_2(1-P_2)}, \tag{10}$$

$$Th_2^{qs} = \sum_{j=1}^{C}\sum_{i=0}^{\alpha} j\mu_2 \cdot \widehat{\pi}(i,j). \tag{11}$$

De la misma manera para ISS, se puede aproximar la distribución de probabilidades de estado estacionario del sistema utilizando (8) en dos fases. La primera fase es calcular, $\pi_1(i)$ como la probabilidad de encontrar i llamadas RT en curso en un sistema $M/M/N_{rt}/N_{rt}$ únicamente con tráfico RT. La segunda fase consiste en calcular $\widehat{\pi}_2(j|i)$ como la probabilidad estacionaria de encontrar j flujos NRT en un sistema $M/M/1/N-PS$ únicamente con tráfico NRT, i.e., utilizando (4) para cada estado i, las probabilidades de estado estacionario están dadas por:

$$\widehat{\pi}_2(j|i) = \pi_n = {a_2}^n \pi_0, \tag{12}$$

donde

$$a_2 \equiv a_2^{(i)} = \frac{\lambda_{nrt}}{\mu_{nrt}^{(i)}} \quad ; \quad \pi_0 = \frac{1-a_2}{1-{a_2}^{N_{nrt}+1}}.$$

En el régimen QS, en ISS, la probabilidad de bloqueo de flujos NRT es P_{nrt}^{qs}, el retardo de transferencia promedio de un flujo NRT está dado por $E^{qs}[D_{nrt}]$ y se pueden calcular utilizando la distribución (8) como sigue:

$$P_{nrt}^{qs} = \sum_{k=0}^{N_{rt}} \widehat{\pi}(k, N_{nrt}), \tag{13}$$

$$E^{qs}[X_{nrt}] = \sum_{j=1}^{N_{nrt}}\sum_{i=0}^{N_{rt}} j \cdot \widehat{\pi}(i,j), \tag{14}$$

$$E^{qs}[D_{nrt}] = \frac{E^{qs}[X_{nrt}]}{\lambda_{nrt}(1-Pb_{nrt}^{qs})}. \tag{15}$$

3.2. Aproximación Cuasi-estacionaria Generalizada –GQSA–

En esta sección, se describe GQSA introducida en [1]. Con este nuevo método se pretende mejorar la precisión de QSA y proveer una manera de equilibrar tanto complejidad computacional como precisión. Dado que en [1] se ha considerado un ISS para el análisis, a continuación se detalla la aplicación de GQSA únicamente para el CRS.

De la misma manera que con QSA, los parámetros de rendimiento de SU se pueden aproximar utilizando GQSA pero considerando un conjunto de filas vecinas a la fila i del proceso de PU (eje y en la Figura 1), en lugar de considerar sólo la fila i para el análisis. De este modo, el proceso conjunto PU y SU es más probable que alcance el equilibrio estadístico, dado que la duración de tiempo que los eventos de PU permanecen de forma contínua en un conjunto de estados vecinos i es generalmente mayor que el tiempo de duración al permanecer contínuamente en un solo estado i.

En GQSA se ha introducido un nuevo parámetro llamado radio y que se denota por R, $R \in \{0, 1, 2, \cdots, \lceil \frac{C_1}{2} \rceil\}$. Este parámetro sirve para indicar el número de filas vecinas de la fila i que se considerará en el modelo a analizar. El número de filas (estado i y estados adyacentes a i) en el diagrama de transición de estados, Figura 1, con un radio definido para calcular la aproximación es $2R+1$. Además, se define $\Omega(i)$ como el conjunto de estados compuesto por la fila de estados de i y sus $2R$ filas de estados más cercanas como se indica a continuación:

$$\Omega(i)=\begin{cases} \{(i',j) \in \mathcal{S} : 0 \leq i' \leq 2R\}, & si \quad 0 \leq i < R, \\ \{(i',j) \in \mathcal{S} : i-R \leq i' \leq i+R\}, & si \quad R \leq i \leq C_1 - R, \\ \{(i',j) \in \mathcal{S} : C_1 - 2R \leq i' \leq C_1\}, & si \quad C_1 - R < i \leq C_1. \end{cases}$$

Como el conjunto de estados del proceso PU puede comprender un solo estado i, se considera que QSA es un caso especial de GQSA cuando R=0; mientras que el hecho de considerar un $R = \lceil \frac{C_1}{2} \rceil$ conlleva encontrar la solución exacta.

En la Figura 1, se muestra los elementos, probabilidades y distribuciones estacionarias involucradas en GQSA para un CRS:

- $\pi_1(i)$ es la probabilidad de encontrar i sesiones en curso en un sistema $M/M/C_1/C_1$ únicamente con PUs.

- $\widehat{\pi}_2(j|i)$ es la probabilidad estacionaria de encontrar j sesiones en curso en un sistema $M/M/C-i/C-i$ únicamente con SUs.
- $\widehat{\pi}(i,j) = \pi_1(i) \cdot \widehat{\pi}_2(j|i)$ es la aproximación de la probabilidad de estado estacionario de CRS utilizando QSA.
- $\pi_{\Omega(i)}(i,j)$ es la distribución de probabilidades de estado estacionario del nuevo conjunto de estados definido por R para el análisis del sistema.

Finalmente, la probabilidad de estado estacionario (i,j) utilizando GQSA se define como:

$$\pi(i,j) \approx \overline{\pi}(i,j) = \pi_1(i) \cdot \frac{\pi_{\Omega(i)}(i,j)}{\sum_j \pi_{\Omega(i)}(i,j)}. \tag{16}$$

Los valores aproximados de los parámetros para evaluar el rendimiento de SUs en un CRS, se calculan utilizando las probabilidades de estado estacionario definidas en (16) de la siguiente manera:

$$P_2^{gqs} = \sum_{k=c2}^{C} \overline{\pi}(C-k,k), \tag{17}$$

$$P_{ft}^{gqs} = \frac{\lambda_1(P_2 - \overline{\pi}(C_1,C_2))}{\lambda_2(1-P_2)}, \tag{18}$$

$$Th_2^{gqs} = \sum_{j=1}^{C}\sum_{i=0}^{\alpha} j\mu_2 \cdot \overline{\pi}(i,j). \tag{19}$$

En ISS, los parámetros para evaluar el rendimiento de tráfico NRT se calculan utilizando la distribución de probabilidades definidas en (16) de la siguiente manera:

$$P_{nrt}^{gqs} = \sum_{k=0}^{N_{rt}} \overline{\pi}(k, N_{nrt}), \tag{20}$$

$$E^{gqs}[X_{nrt}] = \sum_{j=1}^{N_{nrt}}\sum_{i=0}^{N_{rt}} j \cdot \overline{\pi}(i,j), \tag{21}$$

$$E^{gqs}[D_{nrt}] = \frac{E^{qs}[X_{nrt}]}{\lambda_{nrt}(1-Pb_{nrt}^{qs})}. \tag{22}$$

4. Evaluación numérica y resultados

En esta sección, se estudia el comportamiento de GQSA cuando la separación de escalas temporales varía desde el régimen QS al regimen fluido. Se proveen y analizan los resultados numéricos de la precisión de la aproximación en el cálculo de parámetros para evaluar el rendimiento tanto del CRS como del ISS.

Como punto de partida para este estudio, se ha implementado la solución exacta de los sistemas, utilizando modelos de Markov bi-dimensionales (ver Figura 1 para CRS, y Figura 2 para ISS) con el objetivo de calcular los valores exactos de sus correspondientes parámetros de rendimiento. Luego se ha implementado GQSA, i.e. se aplica la distribución definida en (16) con la finalidad de estimar los parámetros para evaluar el rendimiento de cada sistema.

En el presente trabajo nos centramos en evaluar el error relativo (e_r) de cada parámetro. Por ejemplo, el error relativo de la probabilidad de bloqueo de SUs en un CRS, $e_r(P_2)$, se calcula de la siguiente manera:

$$e_r(P_2) = \frac{|P_2^E - P_2^G|}{P_2^E} \tag{23}$$

donde P_2^E es el valor exacto de la probabilidad de bloqueo de SUs y P_2^G es el valor aproximado de la probabilidad de bloqueo, calculada utilizando (16).

Para evaluar la bondad de GQSA, en términos de complejidad computacional, tiempo de ejecución y precisión, se han estudiado sistemas con diferentes tamaños (número de canales) y diferentes condiciones de carga.

Para configurar las condiciones de carga, se ha procedido de la siguiente manera: primeramente se ha configurado las tasas de servicio a 1, posteriormente se ha ajustado las tasas de llegada para obtener dos condiciones de carga: baja (L) y alta (H), las cuales corresponden a probabilidades de bloqueo de $1 \cdot 10^{-3}$ y $5 \cdot 10^{-2}$, respectivamente.

Dado que utilizamos modelos de Markov bi-dimensionales, en total se considera cuatro configuraciones de carga para cada sistema:

LL condición de baja carga en PUs (tráfico RT), y condición de baja carga en SUs (tráfico NRT).

LH condición de baja carga en PUs (tráfico RT), y condición de alta carga en SUs (tráfico NRT).

HL condición de alta carga en PUs (tráfico RT), y condición de baja carga en SUs (tráfico NRT).

HH condición de alta carga en PUs (tráfico RT), y condición de alta carga en SUs (tráfico NRT).

Con las tasas de llegada ajustadas a la carga especificada (LL, LH, HL o HH), se utiliza un factor de aceleración de eventos f, $10^{-5} \leq f \leq 10^{5}$, para acelerar o retardar la llegada o salida de eventos (de los PUs, en el caso de CRS o del tráfico RT, en el caso de ISS) mientras se mantiene el tráfico ofrecido constante.

Por ejemplo, en un CRS, con la finalidad de analizar los métodos de aproximación cuando la separación de escalas temporales varía desde el régimen QS hacia el régimen fluido, para cada valor de f, la tasa de llegadas de PU y sus correspondientes tasas de servicio se obtienen como $\lambda_1(f) = f \cdot \lambda_1$ y $\mu_1(f) = f \cdot \mu_1$.

En CRS se analiza, para los SUs, la probabilidad de bloqueo, la probabilidad de terminación forzada y el throughput; se consideran los siguientes valores para el número de canales primarios: $C_1 = \{30, 40, 50, 60, 70, 80, 90\}$ y para cada uno de ellos, son considerados los siguientes valores para el número de canales secundarios: $C_2 = \{C_1, (C_1/2), (C_1/5), (C_1/10)\}$.

En ISS se analiza la probabilidad de bloqueo y el retardo de transferencia promedio de tráfico NRT; manteniendo c y L constantes se consideran valores de capacidad total de enlace de $C = \{1{,}92, 3{,}84, 7{,}68\}$ Mbps.

4.1. Evolución de la precisión de GQSA al variar la separación de escalas temporales

Se ha variado el factor de aceleración de eventos f para analizar el comportamiento de las aproximaciones como una función de la separación de escalas temporales. Los resultados se muestran en las Figuras 3 — 10; de ellas se pueden hacer las siguientes observaciones:

1. Los valores de los parámetros de rendimiento obtenidos mediante GQSA alcanzan el valor exacto cuando R es incrementado hasta $C_1/2$ en CRS o $N_{rt}/2$ para ISS.

2. Como se esperaba, cuando el factor de aceleración de eventos f decrece, las curvas tienden al regimen QS para todos los parámetros medidos para evaluar el rendimiento tanto de ISS como de CRS.

3. En ISS, para todos los valores del factor de aceleración f y para todos los parámetros de rendimiento, cuando R aumenta desde 0 hasta $N_{rt}/2$ la aproximaciones se acercan poco a poco al valor exacto. La Figura 3 muestra este comportamiento para la P_{nrt} con una condición de carga LH. En la Figura 4 se puede observar como el $e_r(P_{nrt})$ disminuye gradualmente a medida que el radio se incrementa en GQSA para la evaluación del rendimiento de ISS. Vemos que cada incremento del radio supone una mejora en comparación a utilizar QSA (R=0).

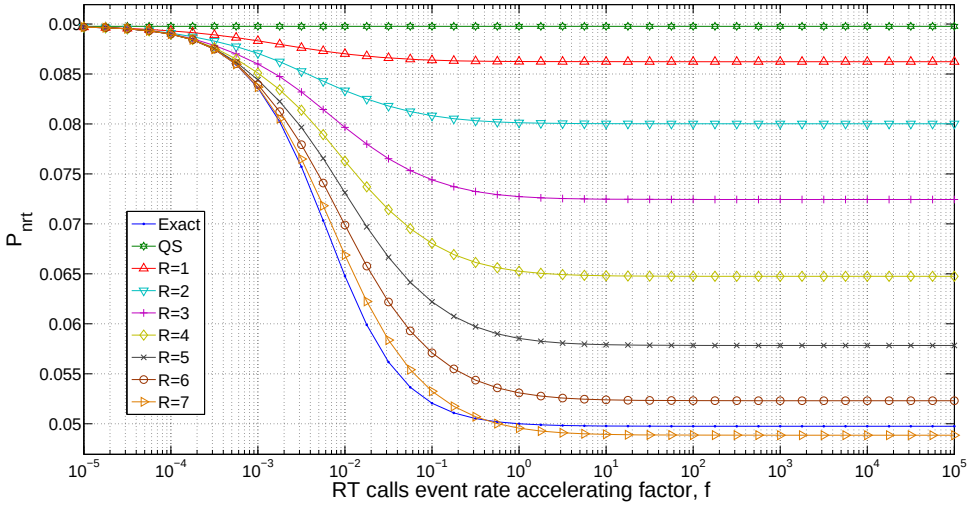

Figura 3: ISS, probabilidad de bloqueo tráfico NRT, condición de carga LH; $\lambda_{rt} =$ 10,812, $\mu_{rt} = 1$, $N_{rt} = 22$; $\lambda_{nrt} = 0{,}317$, $N_{nrt} = 30$; $C = 1{,}92$ Mbps, $c = 64$ kbps, $L = 4$ Mb.

En contraste, como se observa en la Figura 5, en un CRS para todos los valores de f, las curvas correspondientes a $R = 1, \ldots, C_1/2$ no están entre la curva de QSA ($R = 0$) y la curva de valores exactos. En otras palabras, QSA ($R = 0$) sobreestima el valor exacto de P_2 mientras que GQSA ($R > 0$) lo subestima. Sin embargo, este comportamiento no se mantiene en todas las diferentes configuraciones (escenarios considerados) y analizando otros parámetros de rendimiento (ver Figura 6 con $R =$ 1).

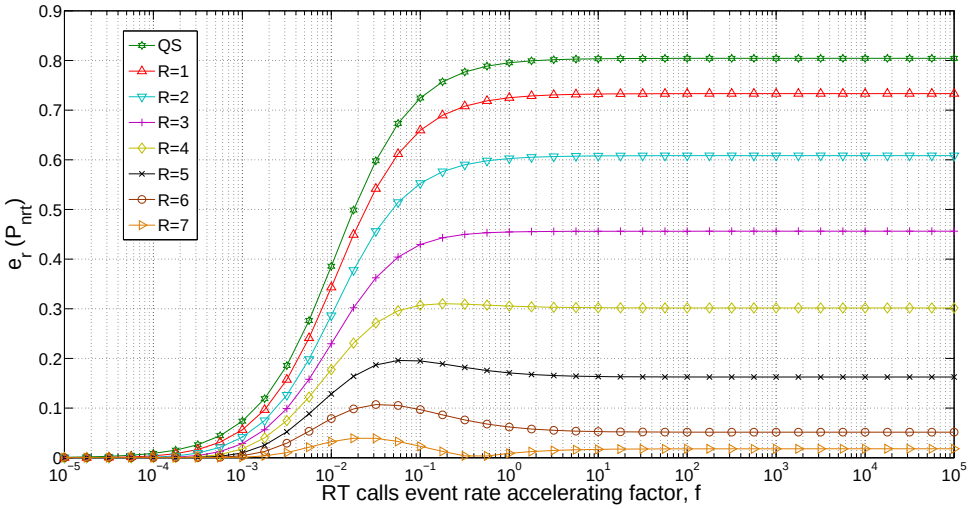

Figura 4: ISS, $e_r(P_{nrt})$, condición de carga LH; $\lambda_{rt} = 10{,}812$, $\mu_{rt} = 1$, $N_{rt} = 22$; $\lambda_{nrt} = 0{,}317$, $N_{nrt} = 30$; $C = 1{,}92$ Mbps, $c = 64$ kbps, $L = 4$ Mb.

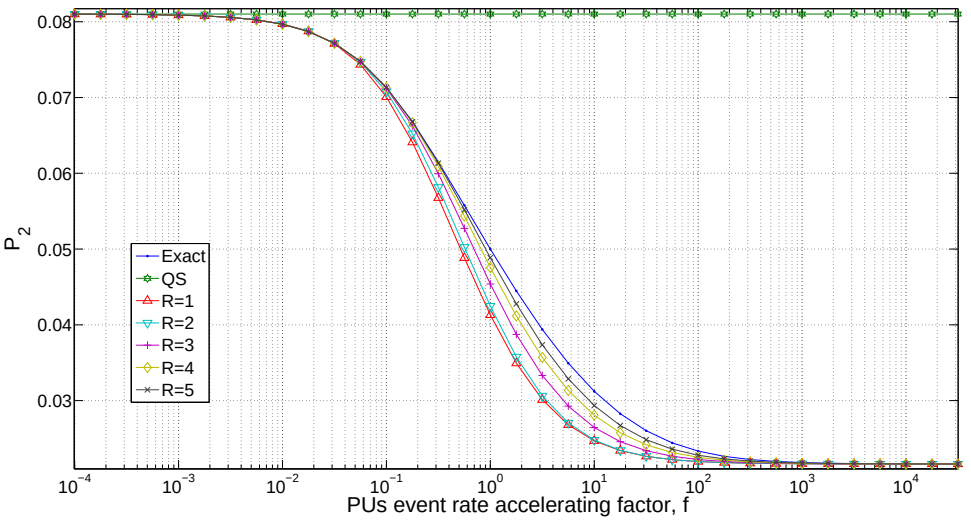

Figura 5: CRS, probabilidad de bloqueo SUs, condición de carga HH; $\lambda_1 = 34{,}596$, $\mu_1 = 1$, $C_1 = 40$; $\lambda_2 = 5{,}455$, $\mu_2 = 1$, $C_2 = 4$.

4. Las Figuras 6 y 7 muestran que para lograr una alta precisión en algunas configuraciones de los sistemas, sólo es necesario incrementar el radio ligeramente ($R = 1$ en la Figura 6, y $R = 3$ en la Figura 7).

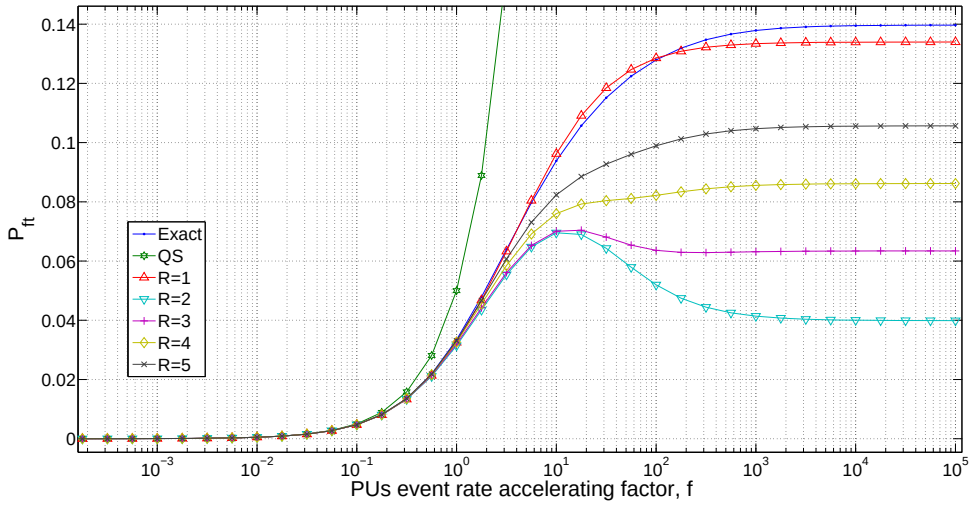

Figura 6: CRS, probabilidad de terminación forzada de SUs, condición de carga HH; $\lambda_1 = 34{,}596$, $\mu_1 = 1$, $C_1 = 40$; $\lambda_2 = 42{,}182$, $\mu_2 = 1$, $C_2 = 40$.

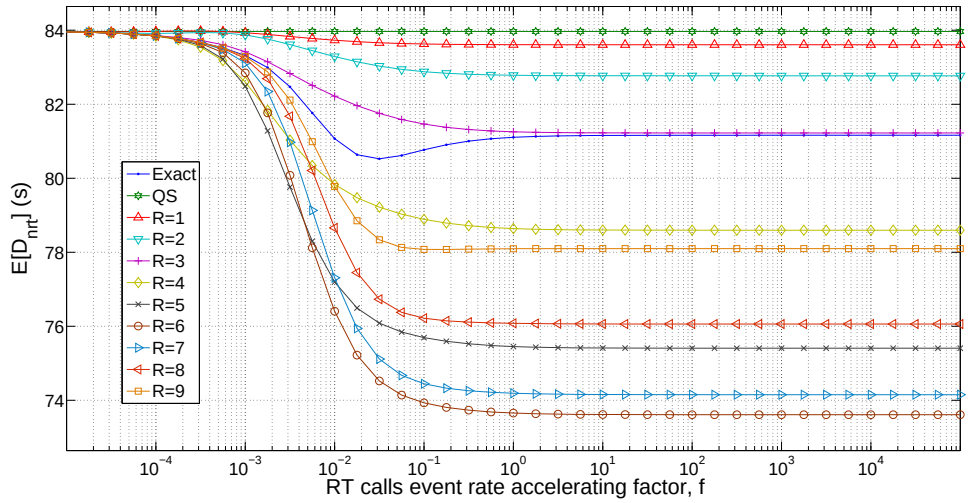

Figura 7: ISS, retardo promedio de transferencia de flujo NRT, condición de carga HH; $\lambda_{rt} = 17{,}132$, $\mu_{rt} = 1$, $N_{rt} = 22$; $\lambda_{nrt} = 0{,}227$, $N_{nrt} = 30$; $C = 1{,}92$Mbps, $c = 64$kbps, $L = 4$Mb.

Se debe tener muy en cuenta que el hecho de incrementar el radio, no siempre asegura una convergencia gradual y monótona al valor exacto. Como se puede observar en la Figura 8 al analizar el error relativo del retardo de transferencia promedio de flujos NRT (ISS); incrementar el radio hasta $R = 3$ mejora la precisión de GQSA. Sin embargo, un incremento de radio desde 4 a 6 hace que la exactitud de GQSA se deteriore. Finalmente, a medida que el radio se va incrementando y toma valores mayores a 6, la precisión de GQSA nuevamente mejora de forma gradual. Es evi-

dente, que el compromiso entre precisión y coste computacional, desalentará el uso de un radio mayor que $R = 3$.

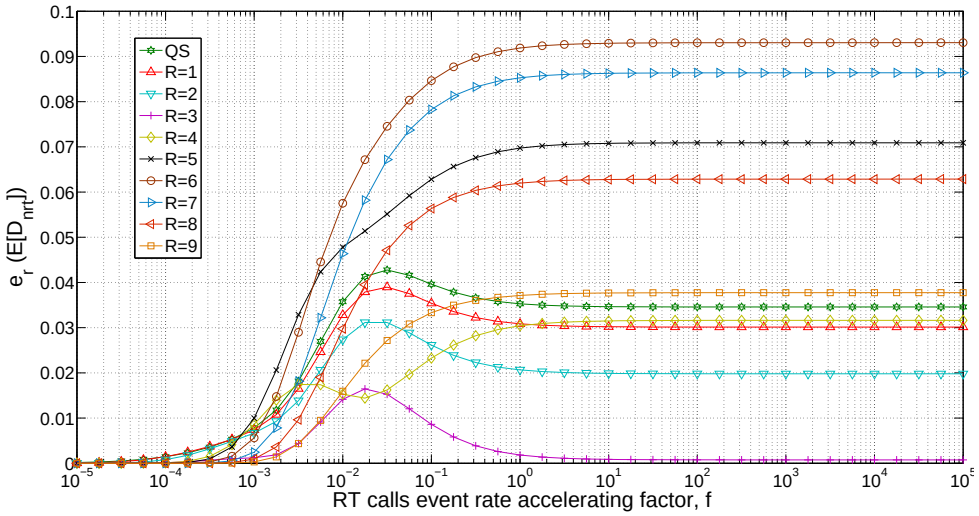

Figura 8: ISS, $e_r(E[D_{nrt}])$, condición de carga HH; $\lambda_{rt} = 17{,}132$, $\mu_{rt} = 1$, $N_{rt} = 22$; $\lambda_{nrt} = 0{,}227$, $N_{nrt} = 30$; $C = 1{,}92$Mbps, $c = 64$kbps, $L = 4$Mb.

Similarmente, en la Figura 9 se observa este comportamiento analizando el error relativo de la probabilidad de terminación forzada de SUs (CRS); donde basta incrementar el radio a 1 para mejorar la precisión con respecto a QSA. El hecho de utilizar radios mayores que 1 no conlleva mayores beneficios en términos de precisión y coste computacional.

5. El comportamiento de GQSA no es monótono en términos de precisión, tanto en CRS como en ISS. Considerando el eje del factor de aceleración de eventos f, la precisión de GQSA comienza siendo buena en el régimen QS; a medida que el factor de aceleración de eventos se va alejando de éste regimen (ver Figura 5 para valores de $10^{-1} \leq f \leq 10^{0}$, y Figura 3 para valores de $10^{-4} \leq f \leq 10^{-2}$), se observa que las curvas de GQSA con radios pequeños comienzan a distanciarse de la curva de valores exactos, i.e. GQSA pierde precisión a medida que nos alejamos de la proximidad al régimen QS. Sorprendentemente, para ciertos valores de R, a medida que f sigue creciendo y se alcanza el régimen fluido ($f > 10^4$), la precisión de GQSA mejora y los valores con ésta (para cualquier valor de $R > 0$) casi superponen a los valores exactos. Este comportamiento es evidente analizando la probabilidad de bloqueo de SUs en el CRS (ver Figura 10) con cualquier tamaño de sistema

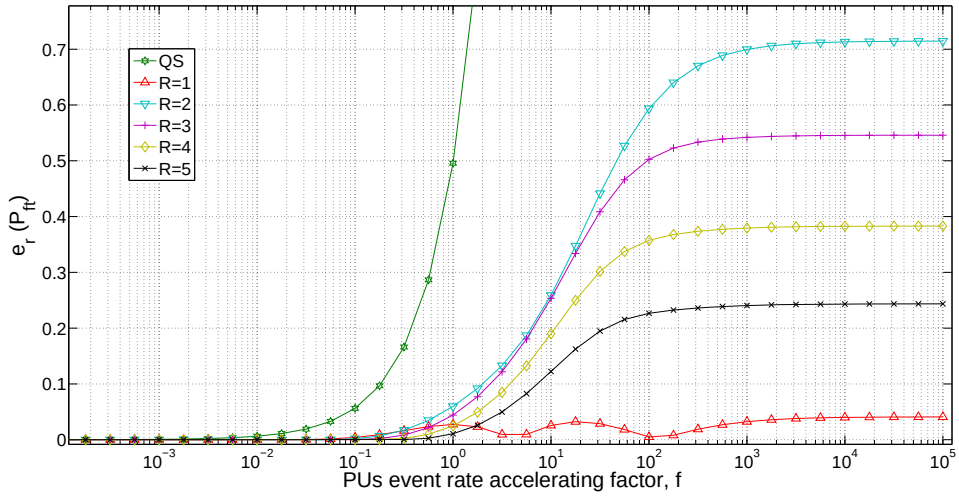

Figura 9: CRS, $e_r(P_{ft})$, condición de carga HH; $\lambda_1 = 34{,}596$, $\mu_1 = 1$, $C_1 = 40$; $\lambda_2 = 42{,}182$, $\mu_2 = 1$, $C_2 = 40$.

y cualquier condición de carga. También se nota este comportamiento en ISS, al analizar el retardo promedio de transferencia de flujos NRT con un radio $R = 3$ y las especificaciones detalladas en la Figura 8.

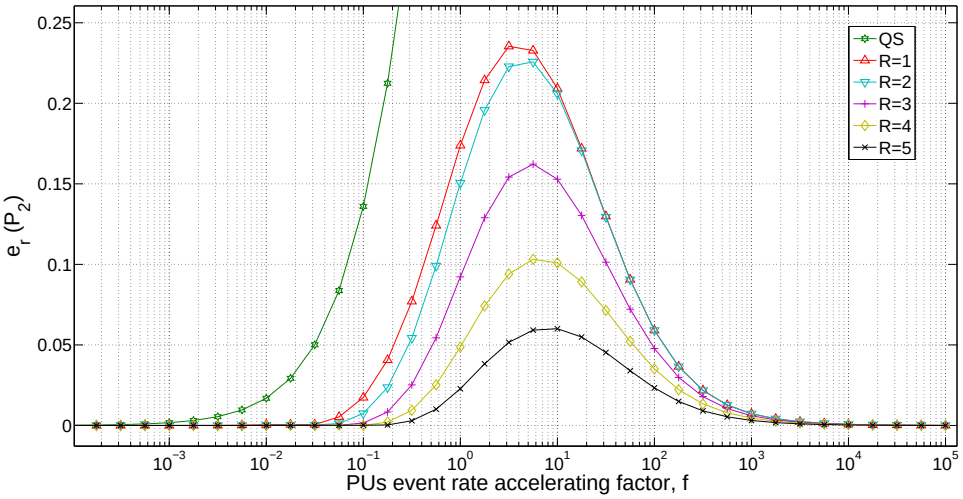

Figura 10: CRS, $e_r(P_2)$, condición de carga HH; $\lambda_1 = 34{,}596$, $\mu_1 = 1$, $C_1 = 40$; $\lambda_2 = 5{,}455$, $\mu_2 = 1$, $C_2 = 4$.

El comportamiento de GQSA detallado en las observaciones 4 y 5, podría ser debido a la forma en que se elige el subconjunto de estados $\Omega(i)$. Se debe tener en cuenta la asimetría con respecto a la fila i, cuando $0 \leq i < R$ y $C_1 - R < i \leq C_1$. Este aspecto

requiere mayor investigación.

4.2. Compromiso entre precisión y coste computacional de los métodos de aproximación

Uno de los resultados a considerar, es el coste computacional al utilizar GQSA en la evaluación del desempeño de los sistemas. Utilizando notación de Landau y considerando que para resolver una cadena de Markov con n estados por eliminación gaussiana, se tiene una complejidad computacional $O(n^3)$ [1]; la complejidad computacional de la solución exacta viene dado por $O(n_E^3)$; donde $n_E = (\frac{C_1}{2} + C_2 + 1) \cdot (C_1 + 1)$ para CRS y $n_E = (N_{rt} + 1) \cdot (N_{nrt} + 1)$ para ISS.

Para hallar la complejidad computacional de la solución aproximada utilizando GQSA, realizamos una sumatoria del resultado de elevar al cubo el número total de estados n_G del conjunto de estados definido por $\Omega(i)$ y el radio R fijado para el análisis.

En la Tabla 1 se reflejan el tiempo de ejecución (ET) y la complejidad computacional (CC) de sistemas CRS para aquellos valores de R en los cuales el tiempo de ejecución no supera el tiempo necesario para obtener la solución exacta.

Tabla 1: Coste computacional utilizando GQSA en CRS

$(C_1, C_2) = (80, 8)$			$(C_1, C_2) = (80, 16)$			$(C_1, C_2) = (80, 40)$			$(C_1, C_2) = (80, 80)$		
R	ET (s)	CC $O(n^3)$	R	ET (s)	CC $O(n^3)$	R	ET (s)	CC $O(n^3)$	R	ET (s)	CC $O(n^3)$
0	0,041	$1{,}60{\cdot}10^7$	0	0,054	$2{,}26{\cdot}10^7$	0	0,094	$5{,}38{\cdot}10^7$	0	0,240	$1{,}60{\cdot}10^8$
1	0,121	$4{,}32{\cdot}10^8$	1	0,124	$6{,}09{\cdot}10^8$	1	0,282	$1{,}45{\cdot}10^9$	1	0,595	$4{,}31{\cdot}10^9$
2	0,255	$2{,}00{\cdot}10^9$	2	0,322	$2{,}81{\cdot}10^9$	2	0,759	$6{,}71{\cdot}10^9$	2	1,902	$1{,}99{\cdot}10^{10}$
3	0,509	$5{,}46{\cdot}10^9$	3	0,727	$7{,}69{\cdot}10^9$	3	1,627	$1{,}84{\cdot}10^{10}$	3	4,036	$5{,}46{\cdot}10^{10}$
4	1,006	$1{,}15{\cdot}10^{10}$	4	1,424	$1{,}63{\cdot}10^{10}$	4	3,354	$3{,}90{\cdot}10^{10}$	4	9,807	$1{,}16{\cdot}10^{11}$
5	1,768	$2{,}09{\cdot}10^{10}$	5	2,367	$2{,}95{\cdot}10^{10}$	5	11,317	$7{,}09{\cdot}10^{10}$	5	13,534	$2{,}11{\cdot}10^{11}$
6	2,668	$3{,}43{\cdot}10^{10}$	6	3,696	$4{,}85{\cdot}10^{10}$	6	9,068	$1{,}17{\cdot}10^{11}$	6	22,445	$3{,}48{\cdot}10^{11}$
Exacto	3,381	$6{,}25{\cdot}10^{10}$	Exacto	5,396	$9{,}84{\cdot}10^{10}$	7	10,943	$1{,}78{\cdot}10^{11}$	7	32,971	$5{,}34{\cdot}10^{11}$
						Exacto	14,320	$2{,}82{\cdot}10^{11}$	8	47,889	$7{,}75{\cdot}10^{11}$
									Exacto	57,286	$9{,}41{\cdot}10^{11}$

En la Figura 11 se representa la CC del cálculo exacto y la de GQSA con el radio máximo obtenido para cada sistema. Se representan sistemas CRS con condición de carga HH, para los cuales, como se puede observar en la Figura 10, el error relativo en la probabilidad de bloqueo de tráfico SUs disminuye con cada incremento de R.

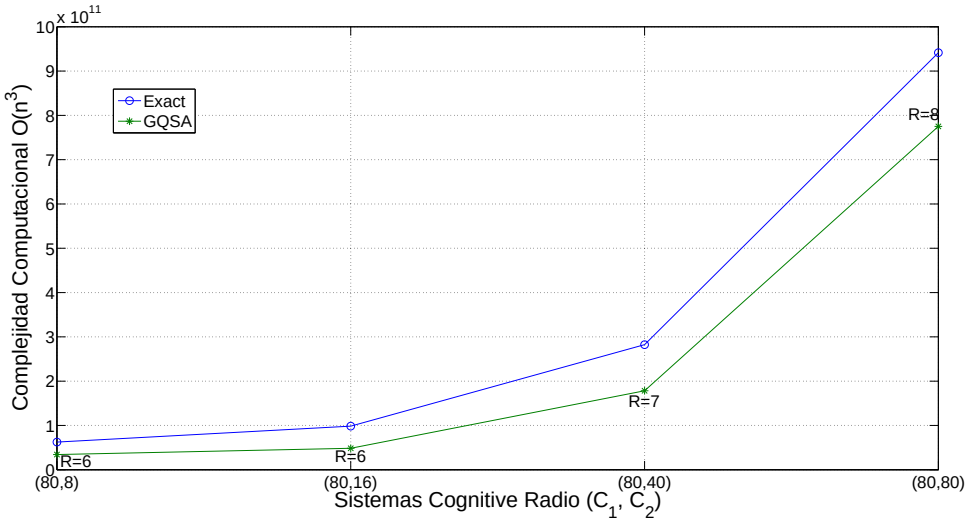

Figura 11: CRS, Complejidad computacional utilizando notación de Landau, condición de carga HH en los sistemas.

De la misma manera para ISS, en la Tabla 2 se reflejan el tiempo de ejecución (ET) y la complejidad computacional (CC) de varios sistemas para aquellos valores de R en los cuales el tiempo de ejecución no supera el tiempo necesario para obtener la solución exacta.

Tabla 2: Coste computacional utilizando GQSA en ISS

$(N_{rt}, N_{nrt}) = (22, 30)$			$(N_{rt}, N_{nrt}) = (44, 60)$			$(N_{rt}, N_{nrt}) = (66, 90)$			$(N_{rt}, N_{nrt}) = (88, 120)$		
R	ET (s)	CC $O(n^3)$	R	ET (s)	CC $O(n^3)$	R	ET (s)	CC $O(n^3)$	R	ET (s)	CC $O(n^3)$
0	0.039	$6.85{\cdot}10^5$	0	0.066	$1.02{\cdot}10^7$	0	0.361	$5.05{\cdot}10^7$	0	4.531	$1.58{\cdot}10^8$
1	0.059	$1.85{\cdot}10^7$	1	0.147	$2.76{\cdot}10^8$	1	0.595	$1.36{\cdot}10^9$	1	5.129	$4.26{\cdot}10^9$
2	0.065	$8.56{\cdot}10^7$	2	0.324	$1.28{\cdot}10^9$	2	1.531	$6.31{\cdot}10^9$	2	8.633	$1.97{\cdot}10^{10}$
3	0.088	$2.35{\cdot}10^8$	3	0.743	$3.50{\cdot}10^9$	3	2.930	$1.73{\cdot}10^{10}$	3	10.242	$5.41{\cdot}10^{10}$
Exacto	0.091	$3.62{\cdot}10^8$	4	1.317	$7.45{\cdot}10^9$	4	5.625	$3.68{\cdot}10^{10}$	4	15.568	$1.15{\cdot}10^{11}$
			5	2.159	$1.36{\cdot}10^{10}$	5	9.650	$6.72{\cdot}10^{10}$	5	23.344	$2.10{\cdot}10^{11}$
			Exacto	3.192	$2.07{\cdot}10^{10}$	6	14.082	$1.11{\cdot}10^{11}$	6	33.918	$3.46{\cdot}10^{11}$
						7	22.260	$1.70{\cdot}10^{11}$	7	53.893	$5.32{\cdot}10^{11}$
						Exacto	23.415	$2.27{\cdot}10^{11}$	8	72.988	$7.75{\cdot}10^{11}$
									Exacto	111.103	$1.25{\cdot}10^{12}$

En la Figura 12 se representa la CC del cálculo exacto y la de GQSA con el radio máximo obtenido para cada sistema. Se representan sistemas ISS con condición de carga LH, para los cuales, como se puede observar en la Figura 4, el error relativo en la probabilidad de bloqueo de N_{nrt} disminuye con cada incremento de R.

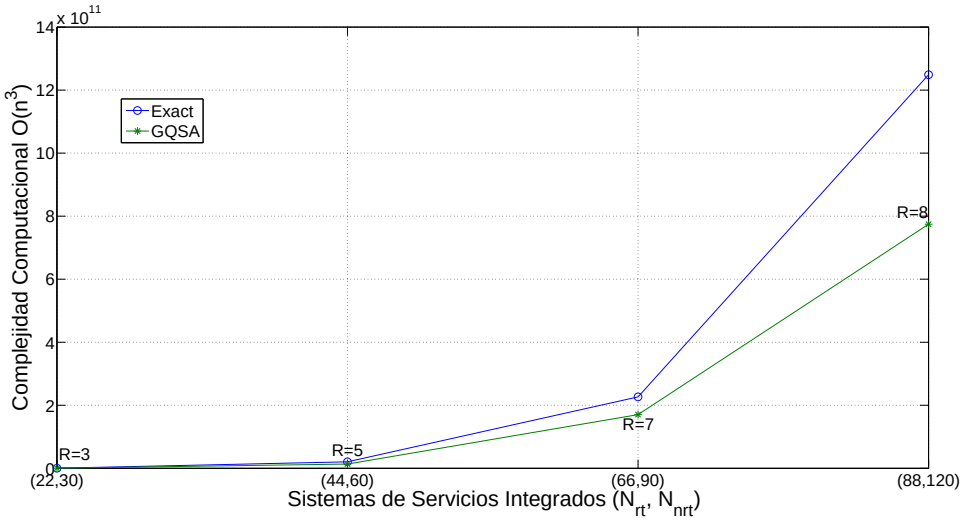

Figura 12: ISS, Complejidad computacional utilizando notación de Landau, condición de carga LH en los sistemas.

Con la finalidad de medir el compromiso entre precisión y costo computacional al utilizar GQSA, en las Figuras 13 y 14 se representa el error relativo e_r versus el tiempo de ejecución para diferentes valores de radio R. Se debe tener en cuenta que $R = 0$ corresponde al caso de QSA. En estas curvas, se ha condigurado el valor de f de tal manera que en régimen QS, el e_r está normalizado a 1 % y 10 % para cada parámetro de rendimiento de los sistemas en análisis. Las Figuras 13 y 14 muestran únicamente los resultados para aquellos valores de R en los cuales el tiempo de ejecución no supera el tiempo necesario para obtener la solución exacta.

Con respecto al comportamiento de GQSA, se constató lo siguiente:

- En ISS, el error relativo decrece cuando el radio utilizado en GQSA se incrementa. Aunque no se representan los resultados en el documento por limitaciones de espacio, se ha observado el mismo comportamiento en todos los parámetros de rendimiento, para todas las condiciones de carga a los que ha sido sometido el sistema y todos los tamaños.

- Se observa un comportamiento bastante diferente, por ejemplo, en CRS con condiciones de carga LL y HL. Analizando las probabilidades de bloqueo y de terminación forzada con condición de carga LL y grandes tamaños de sistemas, se observó que es mejor utilizar QSA que utilizar GQSA con algunos valores para el radio.

- En la Figura 13 para la curva que inicia con $e_r = 1\,\%$, se observa que desde $R = 0$ a $R = 1$, el e_r decrece en un 30 % mientras que el tiempo de ejecución se incrementa en 94 %. Cuando la aproximación inicial es más pobre ($e_r = 10\,\%$ con R=0), la precisión mejora más lentamente a medida que se incrementa R.

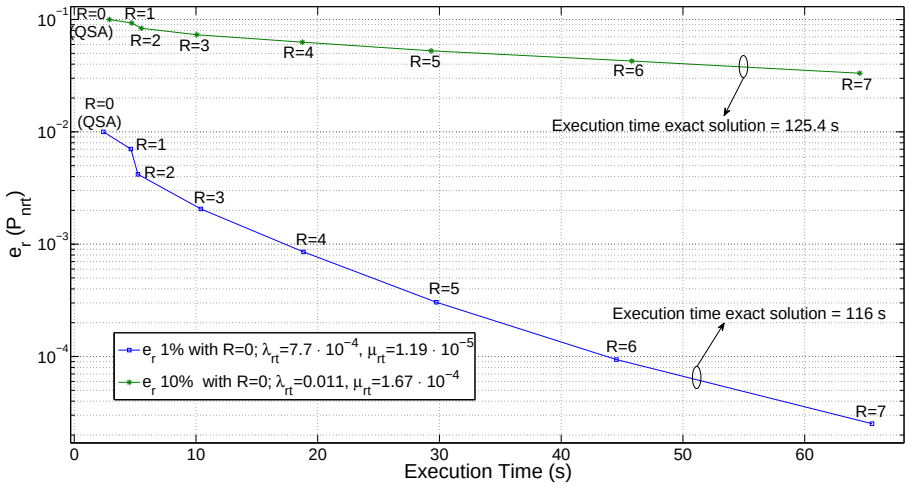

Figura 13: ISS, e_r en probabilidad de bloqueo, condición de carga LL; $C =$ 7,68 Mbps, $c = 64$ kbps, $L = 4$ Mb; $N_{rt} = 88$, $N_{nrt} = 120$

- La Figura 14 muestra que el e_r se incrementa abruptamente desde $R = 0$ a $R = 1$ y después va disminuyendo gradualmente. El mismo comportamiento se observa sin importar cuál es el e_r inicial (1 % o 10 %).

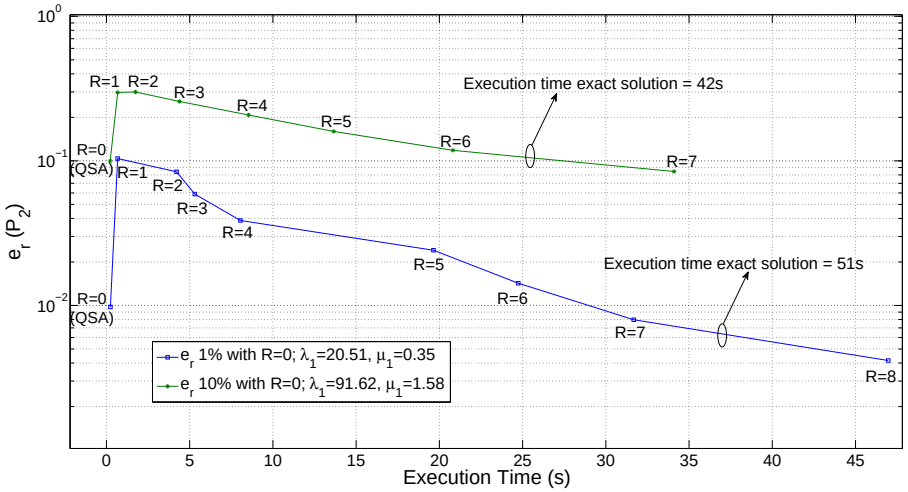

Figura 14: CRS, e_r en probabilidad de bloqueo, condición de carga LL; $C_1 = 80$, $C_2 = 80$; $\lambda_2 = 71{,}62$, $\mu_2 = 1$

5. Conclusiones

En este trabajo se ha estudiado dos métodos de aproximación basados en la descomposición de escalas temporales para el análisis de sistemas cognitive radio y sistemas de servicios integrados que, a nivel de modelo, presentan importantes diferencias cualitativas. Dichos sistemas se han modelado utilizando procesos de Markov en tiempo continuo. Se ha evaluado el comportamiento de las aproximaciones cuando la separación de escalas temporales varía desde el régimen QS hacia el régimen fluido. Se ha medido el compromiso entre precisión y costo computacional. Durante el estudio, se ilustra cómo la generalización de la aproximación cuasi-estacionaria (GQSA) muestra un comportamiento, en términos de precisión, no encontrado previamente en el análisis de otros sistemas.

Los resultados numéricos demuestran que, contrario a lo que se esperaba, el error relativo de los parámetros de rendimiento medidos utilizando GQSA, no siempre decrece cuando se incrementa el radio, i.e. incrementar el radio no siempre mejora la precisión, en algunos casos ésta empeora; por consiguiente, el costo computacional necesario para ganar en precisión puede ser muy alto en comparación a utilizar QSA con la finalidad de evaluar el rendimiento de los sistemas.

El hecho de saber cuando el error relativo decrece, y cuando no, depende de un modo complejo de varios factores. Algunos de ellos (tipo de sistema, condición de carga, tamaños de los sistemas) se han abordado en este trabajo, mientras que otros requieren mayor investigación, debido a que es difícil de predecir en qué casos la precisión se puede mejorar mediante el nuevo método.

Un hallazgo inesperado es que, en algunos casos específicos, GQSA resulta ser una buena aproximación no sólo para el regimen QS, sino también para el régimen fluido, donde la separación de escalas temporales es despreciable.

6. Agradecimientos

A Dios por las oportunidades que me brinda día a día, por bendecir mi vida y la de mi familia.

A mi esposa Karin y mi hija Sammy por todo su amor, apoyo incondicional, comprensión y ánimo en todo momento, gracias por ser pilar fundamental en mi vida y mi inspiración. También a mis padres, hermanos y familia por alentarme y ser mi soporte en toda cisrcunstancia.

Al Dr. Vicent Pla y al Dr. Jorge Martínez por la confianza depositada en mí, por su inigualable paciencia y por el tiempo que han dedicado a este trabajo desde su inicio, contribuyendo con sus ideas y supervisando los resultados para que este trabajo sea cada vez mejor. También agradezco el apoyo a los demás miembros del grupo GIRBA y a mis compañeros de laboratorio.

Gracias a todos.

7. Reconocimiento

El presente trabajo y mis estudios de Máster han sido financiados por la Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) de la República del Ecuador, mediante una beca de estudios contemplada dentro del "Programa de Estudios de Cuarto Nivel de Formación Académica en el Exterior, Convocatoria Abierta 2011" con el auspicio de la Escuela Superior Politécnica de Chimborazo.

Referencias

[1] Y. Huang, K. Ko, and M. Zukerman, "A generalized quasi-stationary approximation for analysis of an integrated service system," *IEEE Communications Letters*, vol. 16, no. 11, pp. 1884–1887, Nov. 2012.

[2] J. Martinez-Bauset, V. Pla, J. Vidal, and L. Guijarro, "Approximate analysis of cognitive radio systems using time-scale separation and its accuracy," *IEEE Communications Letters*, vol. 17, no. 1, pp. 35–38, Jan. 2013.

[3] G. G. Yin and Q. Zhang, *Discrete-time Markov chains: two-time-scale methods and applications*. Springer, 2005, vol. 55.

[4] O. J. Boxma, A. F. Gabor, R. Núñez-Queija, and H.-P. Tan, "Performance analysis of admission control for integrated services with minimum rate guarantees," in *Proceedings of NGI'06*, 2006, pp. 41–47.

[5] F. Hubner and P. Tran-Gia, "Quasi-stationary analysis of a finite capacity asynchronous multiplexer with modulated deterministic input," *ITC-13, Copenhagen*, 1991.

[6] S. Liu and J. Virtamo, "Performance analysis of wireless data systems with a finite population of mobile users," in *Proceedings of the 19th International Teletraffic Congress ITC 19*, 2005, pp. 1295–1304.

[7] I. F. Akyildiz, W. Lee, M. Vuran, and S. Mohanty, "A survey on spectrum management in cognitive radio networks," *IEEE Communications Magazine*, vol. 46, no. 4, pp. 40–48, 2008.

[8] J. W. Roberts, "Internet traffic, qos, and pricing," *Proceedings of the IEEE*, vol. 92, no. 9, pp. 1389–1399, 2004.

[9] H. Al-Mahdi, M. A. Kalil, F. Liers, and A. Mitschele-Thiel, "Increasing spectrum capacity for ad hoc networks using cognitive radios: an analytical model," *IEEE Communications Letters*, vol. 13, no. 9, pp. 676–678, Oct. 2009.

[10] J. Peha, "Sharing spectrum through spectrum policy reform and cognitive radio," *Proceedings of the IEEE*, vol. 97, no. 4, pp. 708–719, 2009.

[11] H. A. Simon and A. Ando, "Aggregation of variables in dynamic systems," *Econometrica: Journal of The Econometric Society*, pp. 111–138, 1961.

[12] P. J. Courtois, *Decomposability: queueing and computer system applications*. Academic Press New York, 1977, vol. 194.

[13] V. Alexiades and A. D. Solomon, *Mathematical modeling of melting and freezing processes*. Taylor & Francis, 1993.

A. Artículos

Approximate Analysis of Wireless Systems Based on Time-Scale Decomposition

Luis Tello-Oquendo, Vicent Pla and Jorge Martinez-Bauset
Dept. of Communications, Universitat Politècnica de València (UPV)
ETSIT, Camí de Vera s/n, 46022 Valencia, Spain
Email: luiteloq@teleco.upv.es, {vpla,jmartinez}@dcom.upv.es

***Abstract*—Markov chains are a widely used modeling tool for wireless communication networks. The system size and the existence of different user types often make the analysis of the Markov chain computationally intractable. When the events of each user type occur at sufficiently separated time scales, the so-called quasi-stationary approximation (QSA) has proven to be accurate and highly efficient. Recently, a generalization of the quasi-stationary approximation (GQSA) has been introduced. The new approximation aims to improve the accuracy at the price of higher computational cost. In this paper, we carry out a comparative study of the accuracy and computational cost of both approximation methods QSA and GQSA. In particular, we explore the evolution of accuracy as the separation between time scales varies, and the trade-off between accuracy and computational cost. Our results indicate that while the new GQSA improves the accuracy in some instances, it does not occur in all of them; and more importantly, it is difficult to predict in which cases accuracy can be enhanced by the new method.**

***Index Terms*—Wireless systems, cognitive radio systems, integrated services systems, traffic analysis, quasi-stationary approximation, time-scale decomposition.**

I. Introduction

Continuous-time Markov chains (CTMC) are commonly used for modeling communication systems in order to study their performance. However, when the size of the systems is large, the computational cost to calculate their performance is greatly increased. Therefore, it is very necessary develop various approximations techniques to reduce the computational cost. One of these computationally efficient approximations, based in time-scale decomposition and often used is the quasi-stationary approximation (QSA) [1]–[3].

In [4] the authors introduce a new method, also based in time-scale decomposition, called Generalized Quasi-Stationary Approximation (GQSA), that provides a way to trade off computational complexity and accuracy. They apply it to an integrated services system (ISS) that serve short-lived non-real-time and long-lived real-time traffic.

Using the newly introduced method, we assess its behavior in a Cognitive Radio System (CRS), which at the model level present qualitative important differences with respect to the resources available for each type of user and the service rates in each state of the model as is described in Section II. The aim is to explore GQSA in both CRS and ISS.

The Cognitive Radio concept proposes to boost spectrum utilization by allowing cognitive users (secondary users, SU) to access the licensed wireless channel in an opportunistic manner so that interference to licensed users (primary users, PU) is kept to a minimum [5]. Since in CRS there are different types of users, the cardinality of state space increases rapidly with the number of channels, implying the growth of the computational complexity related with the solution of the CTMC associated with the system. For this reason an approximation is tipically required.

On the other hand, in an ISS, future generation broadband networks are expected to support a large variety of applications, typically grouped into two broad categories: real-time (RT) (e.g. voice and video) and non-real-time (NRT) (e.g. web-browsing, email and file-transfer) [6]. When the number of channels is large the computational complexity of solving the CTMC associated with the system becomes prohibitive. Therefore, computationally efficient approximations are required [4].

We approach the problem from the traffic perspective and develop two analytical models, one for CRS and another one for ISS, to evaluate the performance of these systems. We consider that the dynamics of the user types (in CRS) or traffic types (in ISS) operates at sufficiently separated time-scales, allowing to use approximation methods based on time-scale decomposition to simplify the computations.

The contribution of this paper is threefold. First, we evaluate the GQSA by applying it to a system (CRS) different from the one studied in [4] (ISS). Second, in both systems (CRS and ISS) we assess the behavior of the approximation when the separation of time-scales vary from the quasi-stationary regime to the fluid regime. Third, we analyze the trade-off between accuracy and computational cost of the approximation methods based on time-scale decomposition.

The rest of the paper is structured as follows. In Section II we describe the Markov models and detail the characteristics of the systems. Section III presents QSA and GQSA approximations based on time-scale decomposition to simplify the computations. Section IV detail the numeric evaluation and show the results of performance metrics of the systems to validate the accuracy and computational cost of approximations. Finally, the conclusions are presented in Section V.

II. Description of Models and Exact Analysis

In this section, we detail the characteristics of the systems and describe the CTMC models associated with them.

1) Cognitive Radio System : As in [7], we model the PU and SU traffic at the session (connection) level and ignore interactions at the packet level (scheduling, buffer management, etc.). We assume an ideal MAC layer for SUs, which allows a perfect sharing of the allocated channels among the active SUs (all active SUs get the same bandwidth portion), introduce zero delay and whose control mechanisms consume zero resources. In addition, we also assume that an active SU can sense

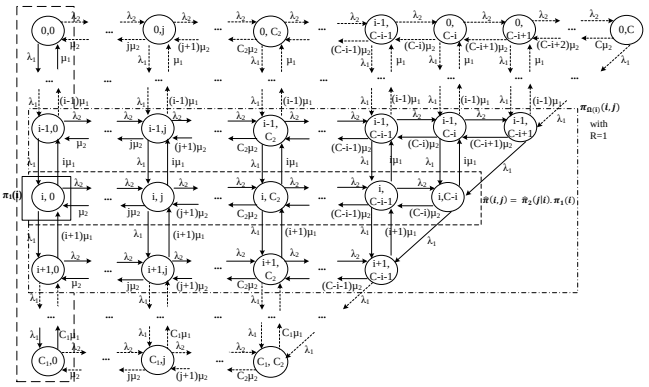

Figure 1. State-transition diagram, Cognitive Radio System.

the arrival of a PU in the same channel instantaneously and reliably. In this sense, the performance parameters obtained can be considered as an upper bound.

The Cognitive Radio System has C_1 *primary channels* (PCs) that can be shared by PUs and SUs, and C_2 *secondary channels* (SCs) only for SUs. Let $C = C_1 + C_2$ be the total number of channels in the system. Note that the SCs can be obtained from e.g. unlicensed bands, as proposed in [8]. This assumption is applicable to the *coexistence* deployment scenario for CRNs [9]. Alternatively, as it might be of commercial interest for the primary and secondary networks to *cooperate*, the secondary channels may be obtained based on an agreement with the primary network [9].

A SU in the PCs might be forced to vacate its channel if a PU claims it to initiate a new session. As SUs support *spectrum handover*, a vacated SU can continue with its ongoing communication if a free channel is available. Otherwise, it is *forced to terminate*.

For the sake of mathematical tractability, Poisson arrivals and exponentially distributed service times are assumed. The arrival rate for PU (SU) sessions is λ_1 (λ_2), their service rate is μ_1 (μ_2), and requests consume 1 (1) channel when accepted.

We denote by (i, j) the system state, when there are i ongoing PU sessions and j SU sessions. The set of feasible states is $\mathcal{S} := \{(i, j) : 0 \leq i \leq C_1, 0 \leq i + j \leq C\}$ and the cardinality of $\mathcal{S}$ is $|\mathcal{S}| = (\frac{C_1}{2} + C_2 + 1) \cdot (C_1 + 1)$. The state-transition diagram of the system is depicted in Fig. 1.

Given the set of feasible states and their transitions in a CTMC, we can construct the global balance equations and the normalization equation. From these we calculate the steady-state probabilities denoted as $\pi(i, j)$.

The system performance parameters are determined as follows,

$$P_1 = \sum_{k=0}^{C_2} \pi(C_1, k) \quad , \quad P_2 = \sum_{k=C_2}^{C} \pi(C - k, k), \tag{1}$$

$$P_{ft} = \frac{\lambda_1(P_2 - \pi(C_1, C_2))}{\lambda_2(1 - P_2)}, \tag{2}$$

$$Th_2 = \sum_{j=1}^{C} \sum_{i=0}^{\alpha} j\mu_2 \cdot \pi(i, j), \tag{3}$$

where P_1 is the PUs blocking probability, which clearly coincides with the one obtained in an Erlang-B loss model with C_1 servers; P_2 is the SUs blocking probability, i.e. the fraction of SU sessions rejected upon arrival as they find the system full; P_{ft} is the forced termination probability of the SUs, i.e. the rate of SU sessions forced to terminate divided by the rate of accepted SU sessions; Th_2 is the SUs throughput, i.e the rate of SU sessions successfully completed and $\alpha = \min(C_1, C - j)$.

2) Integrated Services System: We use the same model defined in [4] for an Integrated Services System that serve real-time (RT) and non-real-time (NRT) traffic. We consider a link whose limited resources (C Mbps in total) are shared amongst RT and NRT requests. The RT traffic is given strict priority over the NRT traffic. We initially assume that all RT calls are of the same class each requiring one channel of rate c b/s during its entire service duration to meet its required QoS. Denote N_{rt} the maximum number of channels for RT calls. When an RT call arrives, it occupies 1 channel if available; otherwise, it is blocked. We set N_{rt} , such that $N_{rt}c$ is sufficiently smaller than C to avoid starvation of the NRT traffic. Let $n_{rt}(t)$ be the number of RT calls in the system at time t, $t \geq 0$, so $\{n_{rt}(t), t \geq 0\}$ is the RT process. NRT flows are served evenly by the leftover capacity from the RT traffic according to the processor sharing (PS) discipline. Let $n_{nrt}(t)$ be the number of NRT flows in the system at time t, $t \geq 0$. Then, $\{(n_{rt}(t), n_{nrt}(t)), t \geq 0\}$ is the joint RT and NRT process. The capacity available for all the NRT traffic at time t is given by $C_{nrt}(t) = C - n_{rt}(t) \cdot c$. The bit-rate of each admitted NRT flow at time t is $c_{nrt}(t) = C_{nrt}(t)/n_{nrt}(t)$, which is updated with RT or NRT admitted arrivals or departures. To satisfy the QoS of admitted NRT flows, the maximum number of concurrent NRT flows is limited to N_{nrt} . Accordingly, an NRT flow arriving at time t is blocked if $n_{nrt}(t) = N_{nrt}$.

We assume Poisson arrivals for RT and NRT requests with rates λ_{rt} and λ_{nrt} respectively. The service time of each admitted RT request is exponentially distributed, its service rate is μ_{rt}. On the other hand, as data sessions generate NRT traffic, their sojourn time will depend on the available resources. The size of the flows generated by the data sessions are exponentially distributed with mean L (bits).

We denote by (i, j) the system state, when there are i ongoing RT calls and j NRT flows. Let $\mathcal{S}$ be the set of feasible states as $\mathcal{S} := \{(i, j) : 0 \leq i \leq N_{rt}, 0 \leq i + j \leq N_{rt} + N_{nrt}\}$ and the cardinality of $\mathcal{S}$ is $|\mathcal{S}| = (N_{rt} + 1)(N_{nrt} + 1)$. The state-transition diagram of the system is depicted in Fig. 2.

Given the set of feasible states and their transitions in a CTMC, we can construct the global balance equations and the normalization equation. From these we calculate the steady-state probabilities denoted as $\pi(i, j)$. We must consider that the service rate of NRT flows varies according to the n_{rt} RT calls in the system as follow:

$$\mu_{nrt}^{(i)} = \frac{C - i \cdot c}{L}. \tag{4}$$

The system performance parameters can be developed as follows:

$$P_{nrt} = \sum_{k=0}^{N_{rt}} \pi(k, N_{nrt}), \tag{5}$$

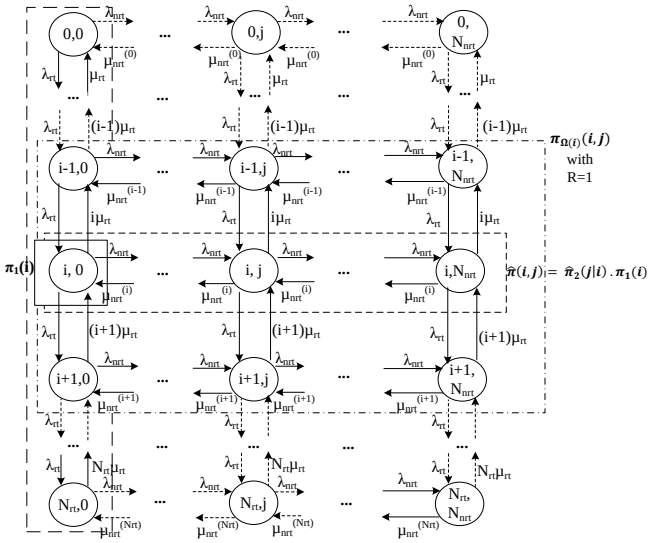

Figure 2. State-transition diagram, Integrated Services System.

$$E\left[X_{nrt}\right] = \sum_{j=1}^{N_{nrt}} \sum_{i=0}^{N_{rt}} j \cdot \pi(i,j), \tag{6}$$

$$E\left[D_{nrt}\right] = \frac{E\left[X_{nrt}\right]}{\lambda_{nrt}(1 - Pb_{nrt})}, \tag{7}$$

where P_{nrt} is the NRT flow blocking probability, $E\left[X_{nrt}\right]$ is the mean number of NRT flows in the system and $E\left[D_{nrt}\right]$ is the NRT flow average transfer delay.

III. Approximations

In terms of the modeling approach, while Markovian models have been developed for the exact analysis of CRS or ISS, they can be numerically cumbersome [3].

This study can be problematic as the higher dimensionality may render the exact analysis computationally intractable. However, when the dynamics of different dimensions (as type of users –PU or SU– in CRS or types or traffic –RT or NRT– in ISS) of the system in analysis operates at sufficiently separated time-scales, one can resort to highly efficient approximations based on time-scale decomposition, which can greatly simplify the computations [7]. In this section, we describe a quasi-stationary approximation and a generalized quasi-stationary approximation to estimate performance parameters in both CRS and ISS.

A. Quasi-stationary Approximation

The simplest approximation, producing easily computable results is the so called quasi-stationary (or, quasi-static) approximation [10].

For instance, in CRS, the approximation can be applied by decoupling the two types of users to analyze their performances. We represent each type of CR user in one dimension of our model. In Fig. 1 we can see PUs process in the y axis and SUs process in the x axis. Since PUs utilize channel resources regardless of the existence of the SUs, the performance analysis for PUs can be evaluated independently in an exact manner. Then the performance of SU can be approximated.

The approximation procedure consists of two stages. The first stage yields the distribution of steady-state probabilities for the PU, i.e. $\{\pi_1(i) : i = 0, ..., C_1\}$, where $\pi_1(i)$ is the probability of finding i ongoing sessions in an $M/M/C_1/C_1$ system with only PUs.

The second stage is to calculate the j SU session distribution of steady-state probabilities by conditioning on the state distribution for the PU: $\widehat{\pi}_2\left(j|i\right)$. Also, $\widehat{\pi}_2\left(j|i\right)$ is the stationary probability of finding j ongoing sessions in an $M/M/C-i/C-i$ system with only SUs.

Both $\pi_1\left(i\right)$ and $\widehat{\pi}_2\left(j|i\right)$ can be determined independently using simple recursions, since their corresponding CTMC are one-dimensional birth-and-death processes.

Finally, the state probability distribution of CRS can be approximated as:

$$\pi(i,j) \approx \widehat{\pi}(i,j) = \pi_1(i) \cdot \widehat{\pi}_2(j|i). \tag{8}$$

Sufficient accuracy can be obtained by using the QS approximation. At the QS regime, the PUs blocking probability is $P_1^{qs} = \pi_1\left(C_1\right)$; the SUs blocking probability is P_2^{qs}, the SUs throughput is Th_2^{qs} and the SUs forced termination probability is P_{ft}^{qs} and can be determined employing distribution (8) as follows [7]:

$$P_2^{qs} = \sum_{k=c2}^{C} \widehat{\pi}(C-k,k), \tag{9}$$

$$P_{ft}^{qs} = \frac{\lambda_1(P_2 - \widehat{\pi}(C_1,C_2))}{\lambda_2(1-P_2)}, \tag{10}$$

$$Th_2^{qs} = \sum_{j=1}^{C} \sum_{i=0}^{\alpha} j\mu_2 \cdot \widehat{\pi}(i,j). \tag{11}$$

In the same way for ISS system, we can approximate the distribution of steady-state probabilities of the system using (8) in two stages: First stage, calculate $\pi_1(i)$ as the probability of finding i ongoing RT calls in an $M/M/N_{rt}/N_{rt}$ system with only RT traffic. The second stage is to calculate $\widehat{\pi}_2\left(j|i\right)$ as the stationary probability of finding j NRT flows in an $M/M/1/N-PS$ system with only NRT traffic, i.e., using (4) for each i, the steady-state probabilities are given by:

$$\widehat{\pi}_2\left(j|i\right) = \pi_n = {a_2}^n \pi_0, \tag{12}$$

where

$$a_2 \equiv a_2^{(i)} = \frac{\lambda_{nrt}}{\mu_{nrt}^{(i)}} \quad ; \quad \pi_0 = \frac{1-a_2}{1-{a_2}^{N_{nrt}+1}}.$$

At the QS regime, in ISS, the NRT flows blocking probability is P_{nrt}^{qs}, NRT flow average transfer delay is given by $E^{qs}\left[D_{nrt}\right]$ and can be determined using distribution (8) as follows:

$$P_{nrt}^{qs} = \sum_{k=0}^{N_{rt}} \widehat{\pi}(k, N_{nrt}), \tag{13}$$

$$E^{qs}\left[X_{nrt}\right] = \sum_{j=1}^{N_{nrt}} \sum_{i=0}^{N_{rt}} j \cdot \widehat{\pi}(i,j), \tag{14}$$

$$E^{qs}\left[D_{nrt}\right] = \frac{E^{qs}\left[X_{nrt}\right]}{\lambda_{nrt}(1 - Pb_{nrt}^{qs})}. \tag{15}$$

B. Generalized Quasi-stationary Approximation

To make this paper self-contained, in this section we describe the GQSA proposed in [4]. For the sake of conciseness, we describe it only for the CRS.

As in QSA, the performance of SU can be approximated using GQSA but considering a set of i neighboring states of the PU process (y axis in Fig. 1) rather than only the state i. In this way, the joint PU and SU process is more likely to reach statistical equilibrium since the time duration that the PU events continuously remains in a set of neighboring i states is generally longer than the time duration that it continuously remains in one i state.

In GQSA a new parameter called radius denoted by R, $R \in \{0, 1, 2, \cdots, \lceil \frac{C_1}{2} \rceil\}$ is introduced, to indicate the size of the set of i neighboring states that we are considering in the model to analyze. The number of rows (state i and adjacent states to i) in the state-transition diagram, Fig. 1, with a defined radius to compute the approximation is $2R + 1$. We define $\Omega(i)$ as the set of states composed of the row of i state and its 2R closest rows of states as follow:

$$\Omega(i) = \begin{cases} \{(i', j) \in \mathcal{S} : 0 \leq i' \leq 2R\}, & 0 \leq i < R, \\ \{(i', j) \in \mathcal{S} : i - R \leq i' \leq i + R\}, & R \leq i \leq C_1 - R, \\ \{(i', j) \in \mathcal{S} : C_1 - 2R \leq i' \leq C_1\}, & C_1 - R < i \leq C_1. \end{cases}$$

As the set of states of PU process can comprise a single PU state, QSA is considered a special case of GQSA when R=0.

In Fig. 1, we show the elements, probabilities and stationary distributions involved in GQSA for a CRS:

- $\pi_1(i)$ is the probability of finding i ongoing sessions in an $M/M/C_1/C_1$ system with only PUs.
- $\widehat{\pi}_2(j|i)$ is the stationary probability of finding j ongoing sessions in an $M/M/C-i/C-i$ system with only SUs.
- $\widehat{\pi}(i,j) = \pi_1(i) \cdot \widehat{\pi}_2(j|i)$ is the approximated steady-state probability of the CRS using QSA.
- $\pi_{\Omega(i)}(i,j)$ is the stationary distribution of the new set of states defined by R for analyze the system.

Finally, the approximated (i,j) steady-state probability using GQSA is defined as follow:

$$\pi(i,j) \approx \overline{\pi}(i,j) = \pi_1(i) \cdot \frac{\pi_{\Omega(i)}(i,j)}{\sum_j \pi_{\Omega(i)}(i,j)}. \tag{16}$$

The approximate values of performance parameters for CRS are calculated using (1), (2) and (3) with the steady-state probabilities defined in (16). Also in ISS, the performance parameters are calculated using (5), (6) and (7) with the steady-state probabilities defined in (16).

IV. Numerical Evaluation and results

In this section, we study the behavior of GQSA when the separation of time-scales vary from QS regime to fluid regime. We provide and discuss numerical results of the accuracy of the approximation in the calculation of performance parameters of CRS and ISS.

As a baseline for our study, we implemented the exact solution of the CTMC system (see Fig. 1 for CRS, and Fig. 2 for ISS) to calculate the exact values of their performance parameters. Then we implemented the GQSA, i.e. we apply the distribution defined in (16) to calculate the performance parameters of each system. We focus on evaluating the relative error (e_r) of each parameter. For instance, the relative error of the blocking probability for SUs in a CRS, $e_r(P_2)$, is computed as

$$e_r(P_2) = \frac{|P_2^E - P_2^G|}{P_2^E} \tag{17}$$

where P_2^E is the exact value of SUs blocking probability and P_2^G is the approximate value of SUs blocking probability calculated using (16).

To evaluate the goodness of GQSA, in terms of computational complexity, execution time and accuracy, we studied the systems with different sizes (number of channels) and different load conditions.

To set the load conditions, we proceed as follows: setting the service rates to 1, we adjusted arrivals rates to obtain two load conditions, low (L) and high (H), which correspond to blocking probabilities $1 \cdot 10^{-3}$ and $5 \cdot 10^{-2}$, respectively.

In total we consider four load configurations for each system:

- LL low load condition for PUs (RT traffic), and low load condition for SUs (NRT traffic).
- LH low load condition for PUs (RT traffic), and high load condition for SUs (NRT traffic).
- HL high load condition for PUs (RT traffic), and Low load condition for SUs (NRT traffic).
- HH high load condition for PUs (RT traffic), and high load condition for SUs (NRT traffic).

With the arrival rates adjusted to the specified load (LL, LH, HL or HH), we use an accelerating factor f, $10^{-5} \leq f \leq 10^5$, to accelerate or decelerate the arrival and departure events (of the PUs, in the case of CRS or of the RT traffic, in the case of ISS) while keeping the offered traffic constant. For instance in a CRS, in order to analyze the approximation methods from the QS regime to the fluid regime, for each value of f, the PU arrival and service rates are obtained as $\lambda_1(f) = f \cdot \lambda_1$ and $\mu_1(f) = f \cdot \mu_1$.

In CRS we analyze blocking probability, forced termination probability and throughput, considering the following values for the number of primary channels: $C_1 = \{30, 40, 50, 60, 70, 80, 90\}$ and for each of them the following values for the number of secondary channels are considering: $C_2 = \{C_1, (C_1/2), (C_1/5), (C_1/10)\}$.

In ISS we analyze blocking probability and average transfer delay for NRT traffic; keeping c and L constant with values for total link capacity of $C = \{1.92, 3.84, 7.68\}$ Mbps.

We varied the accelerating factor f to analyze the behavior of the approximations as a function of the separation of time-

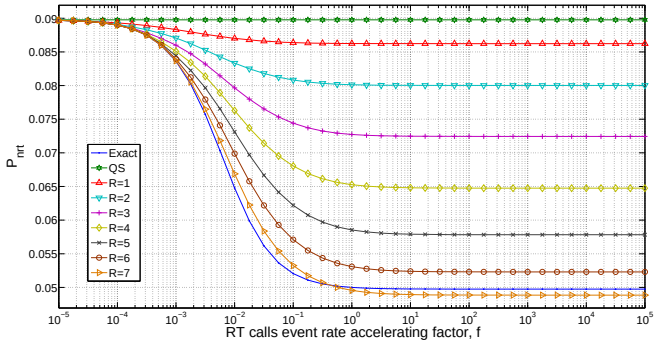

Figure 3. ISS, NRT Blocking Probability, LH load condition; $\lambda_{rt} = 10.812$, $\mu_{rt} = 1$, $N_{rt} = 22$; $\lambda_{nrt} = 0.317$, $N_{nrt} = 30$; $C = 1.92$ Mbps, $c = 64$ kbps, $L = 4$ Mb.

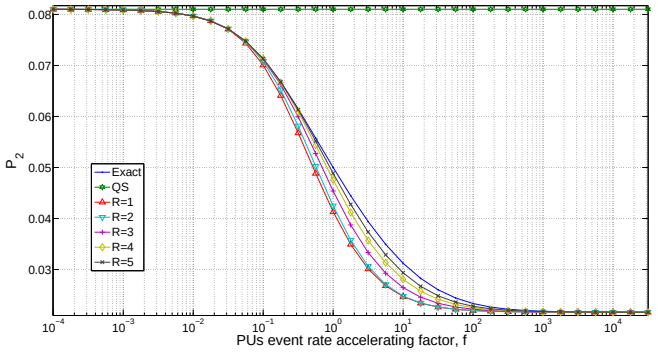

Figure 4. CRS, SUs Blocking Probability, HH load condition; $\lambda_1 = 34.596$, $\mu_1 = 1$, $C_1 = 40$; $\lambda_2 = 5.455$, $\mu_2 = 1$, $C_2 = 4$.

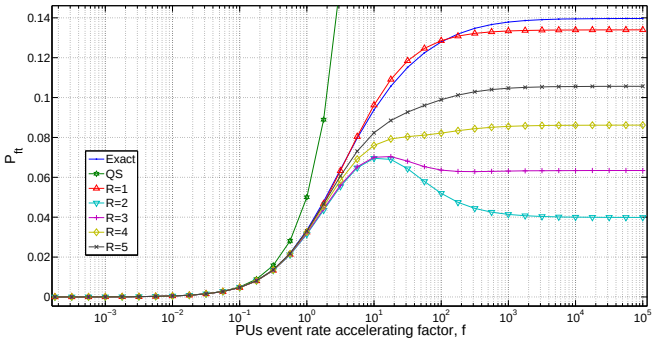

Figure 5. CRS, SUs Forced Termination Probability, HH load condition; $\lambda_1 = 34.596$, $\mu_1 = 1$, $C_1 = 40$; $\lambda_2 = 42.182$, $\mu_2 = 1$, $C_2 = 40$.

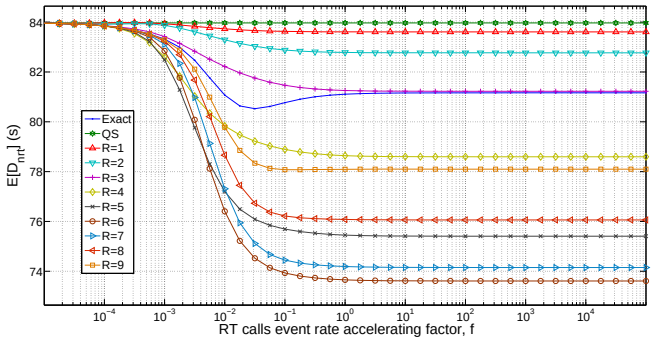

Figure 6. ISS, NRT Flow Average Transfer Delay, HH load condition; $\lambda_{rt} = 17.132$, $\mu_{rt} = 1$, $N_{rt} = 22$; $\lambda_{nrt} = 0.227$, $N_{nrt} = 30$; $C = 1.92$Mbps, $c = 64$kbps, $L = 4$Mb.

scales. The results are shown in Figs. 3–6; from them we can make the following observations:

1) The values of the performance parameter obtained by the GQSA attain the exact value when R is increased to $C_1/2$ in CRS or $N_{rt}/2$ for ISS.
2) As expected, when the accelerating factor f decreases the curves tend to the QS regime for all performance parameters.
3) In ISS, for all values of accelerating factor f and for all performance parameters, when R increases from 0 to $N_{rt}/2$ the approximations approach gradually the exact value. Figure 3 shows this behavior for P_{nrt} with load condition LH.
   In contrast, as can be observed in Fig. 4, in CRS for all values of f, the curves corresponding to $R = 1, \ldots, C_1/2$ are not between the curve for QSA ($R = 0$) and the curve for the exact values. In other words, the QSA ($R = 0$) overestimate the exact value of P_2 whereas the GQSA ($R > 0$) underestimate it. However, this behavior is not maintained for all different configurations (see Fig. 5 with $R = 1$).
4) Figure 5 and 6 show that to achieve a high accuracy in some system configurations, the radius needs only to be increased slightly ($R = 1$ in Fig. 5, and $R = 3$ in Fig. 6).
   Note that increasing the radius not always ensures a gradual and monotonous convergence to the exact value. As can be seen in Fig 6, up to $R = 3$ increasing the radius improved the accuracy of the GQSA. However, increasing further the radius from 4 to 6 the accuracy of the GQSA deteriorates. Finally, as the radius increases beyond 6, the accuracy of the GQSA gradually improves again. Clearly, the trade-off between the accuracy and computational cost will discourage the use of a radius larger than $R = 3$.
5) The behavior of GQSA is not monotonous in terms of accuracy in both ISS and CRS. GQSA accuracy starts being good in QS regime; as the accelerating factor moves away from the QS regime (see Fig. 4 for $10^{-1} \leq f \leq 10^{0}$, and Fig. 3 for $10^{-4} \leq f \leq 10^{-2}$), we observe that the curves of GQSA using small radius begin to distance the curve of the exact values, i.e. GQSA ceases to be accurate. Surprisingly, for certain values of R, as f keeps on growing and we approach the fluid regime ($f > 10^4$), the accuracy of the GQSA improves and the values obtained with it (for any value of $R > 0$) almost overlap the exact ones. This behavior is clearly observed for the blocking probability in CRS with any system size and load condition (see Fig. 4). Also we note this behavior in ISS, for the NRT flow average transfer delay with $R = 3$ and the specifications detailed in Fig. 6.

The behavior in observations 4 and 5, might be due to the

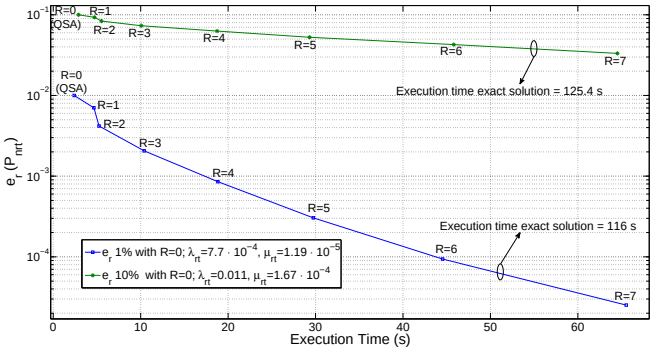

Figure 7. ISS, e_r in Blocking Probability, LL load condition; $C =$ 7.68 Mbps, $c = 64$ kbps, $L = 4$ Mb; $N_{rt} = 88$, $N_{nrt} = 120$

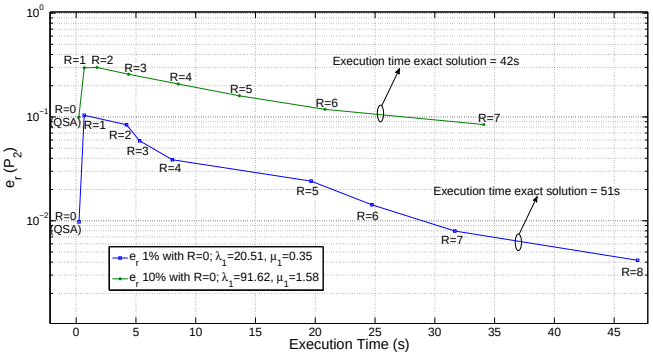

Figure 8. CRS, e_r in Blocking Probability, LL load condition; $C_1 = 80$, $C_2 = 80$; $\lambda_2 = 71.62$, $\mu_2 = 1$

way in which the subset of states $\Omega(i)$ is chosen. Note the asymmetry with respect to row i, when $0 \leq i < R$ and $C_1 - R < i \leq C_1$. This aspect requires further investigation.

In order to measure the trade-off between accuracy and computational cost, in Figs. 7 and 8 we represent the relative error e_r against the execution time for different values of the radius R. Note that $R = 0$ corresponds to the QSA case. In this curves, we have set the value of f so that in QS regime, the e_r is normalized to 1% and 10% for each performance parameter. Figures 7 and 8 show only the results for those values of R in which the execution time that do not exceed the time necessary to obtain the exact solution.

On the behavior of GQSA we note the following:

In ISS, the relative error decreases when the radius of GQSA increases. Although they are not represented here due to the space limitations, we have observed the same behavior in all performance parameters, for all load conditions and system sizes.

A rather different behavior is observed, for instance, in CRS with LL and HL load conditions. Analyzing blocking and forced termination probabilities with LL load condition and large system sizes, is better to use QSA than to use GQSA with some values for the radius.

In Fig. 7 for the curve with initial $e_r = 1\%$, we observe that from $R = 0$ to $R = 1$, e_r decreases in 30% while the execution time increases by 94%. When the initial approximation is poorer ($e_r = 10\%$ with R=0), accuracy improves more slowly as we increase R.

Fig. 8 shows that e_r increases abruptly from $R = 0$ to $R = 1$ and then decline gradually. The same behavior is observed no matter what the initial e_r is (1% or 10%).

V. Conclusions

In this paper we have studied two approximation methods based on time-scale decomposition for the analysis of cognitive radio systems and integrated services systems which, at the model level, present qualitative important differences. We have modeled them as continuous time Markov chains. We assessed the behavior of the approximations when the separation of time-scales vary from the QS regime to the fluid regime. We have measured the trade-off between accuracy and computational cost. During the study, we illustrate how GQSA display a behavior in terms of accuracy, not previously encountered in the analysis of other systems. The numerical results demonstrate, contrary to what one may expect, that the relative error of the performance parameters using GQSA does not always decrease when R is increased, i.e. increasing the radius not always improves the accuracy, in some cases it deteriorates; therefore, the computational cost necessary to gain in accuracy can be very high in comparison to use QSA in order to evaluate the performance of the systems. Knowing when the relative error decreases, and when it does not, depends in a complex way on several factors. Some of them (type of system, load conditions, system sizes) were discussed in the paper while other require further investigation, due to it is difficult to predict in which cases accuracy can be enhanced by the new method. An unexpected finding is that in some specific cases, GQSA is a good approximation not only for QS regime but also in the fluid regime, where the difference in the separation of time-scales is negligible.

References

[1] F. Hubner and P. Tran-Gia, "Quasi-stationary analysis of a finite capacity asynchronous multiplexer with modulated deterministic input," *ITC-13, Copenhagen*, 1991.

[2] S. Liu and J. Virtamo, "Performance analysis of wireless data systems with a finite population of mobile users," in *Proceedings of the 19th International Teletraffic Congress ITC 19*, 2005, pp. 1295–1304.

[3] O. J. Boxma, A. F. Gabor, R. Núñez-Queija, and H.-P. Tan, "Performance analysis of admission control for integrated services with minimum rate guarantees," in *Proceedings of NGI'06*, 2006, pp. 41–47.

[4] Y. Huang, K. Ko, and M. Zukerman, "A generalized quasi-stationary approximation for analysis of an integrated service system," *IEEE Communications Letters*, vol. 16, no. 11, pp. 1884–1887, Nov. 2012.

[5] I. F. Akyildiz, W. Lee, M. Vuran, and S. Mohanty, "A survey on spectrum management in cognitive radio networks," *IEEE Communications Magazine*, vol. 46, no. 4, pp. 40–48, 2008.

[6] J. W. Roberts, "Internet traffic, qos, and pricing," *Proceedings of the IEEE*, vol. 92, no. 9, pp. 1389–1399, 2004.

[7] J. Martinez-Bauset, V. Pla, J. Vidal, and L. Guijarro, "Approximate analysis of cognitive radio systems using time-scale separation and its accuracy," *IEEE Communications Letters*, vol. 17, no. 1, pp. 35–38, Jan. 2013.

[8] H. Al-Mahdi, M. A. Kalil, F. Liers, and A. Mitschele-Thiel, "Increasing spectrum capacity for ad hoc networks using cognitive radios: an analytical model," *IEEE Communications Letters*, vol. 13, no. 9, pp. 676–678, Oct. 2009.

[9] J. Peha, "Sharing spectrum through spectrum policy reform and cognitive radio," *Proceedings of the IEEE*, vol. 97, no. 4, pp. 708–719, 2009.

[10] V. Alexiades and A. D. Solomon, *Mathematical modeling of melting and freezing processes*. Taylor & Francis, 1993.