El papel de la calidad de servicio en las comunicaciones

Autor:

José Luis Poza Luján

Revisores:

José Enrique Simó Ten

Juan Luis Posadas Yagüe

Instituto de Automática e Informática Industrial

(ai2)

Universidad Politécnica de Valencia

(UPV)

Versión: 0.5

Fecha revisión: 10 de noviembre de 2009

Contenidos

Figuras

Ecuaciones

1 Introducción

1.1 Resumen

La calidad de servicio representa actualmente el método de evaluación del funcionamiento de un sistema de comunicaciones, y por extensión, podría ser considerado el método de evaluación de un sistema distribuido.

En el presente documento se realiza una revisión breve, en lo que se trata del concepto de calidad de servicio, y mucho más exhaustiva en el ámbito de los parámetros en los que se plasma el concepto de calidad de servicio. Además de la revisión de parámetros clásicos, se realiza una revisión de los posibles índices que pueden ser empleados como indicadores de calidad de servicio.

1.2 Historial de revisiones

Nº revisión	Fecha	Comentarios
0.0	2007-12	Inicio del documento.
0.1	2008-01	Definiciones de calidad de servicio.
0.2	2008-05	Parámetros básicos de medición de calidad de servicio: revisión bibliográfica.
0.3	2008-06	Ampliación a indicadores empleados en otros sistemas: DCPS.
0.4	2008-10	Análisis de parámetros e indicadores: propuestas.
0.5	2009-07	Revisión global del documento.

1.3 Objetivos del documento

El objetivo del documento es revisar los parámetros más habituales de calidad de servicio para así poder determinar qué indicadores pueden ser empleados en los sistemas distribuidos, con especial incidencia en los sistemas distribuidos de control inteligente.

1.4 Alcance y audiencia

El documento cubre los detalles de los parámetros de calidad de servicio, junto con los indicadores empleados en sistemas actuales de comunicación y control para medir el rendimiento general del sistema. El documento está dirigido a investigadores en comunicaciones y control que deseen tener una visión precisa acerca de los parámetros de calidad de servicio.

1.5 Organización del documento

El documento comienza revisando las definiciones de calidad de servicio, para seguidamente, en el capítulo 3, revisar los parámetros de calidad de servicio comúnmente aceptados como tales. En el capítulo 4 se extiende el concepto de parámetros básicos tratados en el capítulo 3 a parámetros más avanzados que puedan ser adaptados a los sistemas distribuidos de control. Finalmente se exponen algunas conclusiones.

2 La calidad de servicio

2.1 Definición

La calidad se define como un conjunto de propiedades inherentes a algo, que permiten juzgar su valor. Por lo que la calidad de servicio, por tanto, es un concepto que trata de definir los parámetros por los que evaluar un servicio ofrecido. En el campo de los sistemas de comunicaciones, existen diversas definiciones de calidad de servicio, entre las cuales destacan las siguientes.

- Calidad de servicio representa el conjunto de las características tanto cuantitativas como cualitativas de un sistema distribuido necesarias para alcanzar las funcionalidades requeridas por una aplicación [Vogel et al., 1995].
- Conjunto de requisitos del servicio que debe cumplir la red en el transporte de un flujo [Crawley et al., 1998].
- Efecto global de las prestaciones de un servicio que determinan el grado de satisfacción de un usuario al utilizar dicho servicio [ITU, 1994]

Basándose en los conceptos anteriores de calidad de servicio (QoS) en el ámbito de las telecomunicaciones se puede decir que la calidad de servicio se refiere a la capacidad de determinadas redes y servicios para admitir que se fije de antemano las condiciones en que se desarrollarán las comunicaciones. Generalmente, se ha extendido el término a la enumeración de las cualidades mesurables de las redes y servicios de telecomunicaciones que permiten evaluarlos.

Es habitual que los parámetros, por medio de los cuales se definen las calidades de servicio y las políticas que las gestionan, sean específicos a los sistemas en los que se aplica la calidad de servicio. Por ejemplo, en el ámbito de las redes de comunicaciones, la calidad de servicio se entiende como la capacidad de asegurar una tasa de datos en la red (ancho de banda), un retardo o una variación del retardo (habitualmente llamado por su nombre en inglés: jitter), mientras que en otros ámbitos, como el procesamiento distribuido, conceptos como el ancho de banda, son más difíciles de encajar.

Lo que es habitual que sea común en todos los sistemas distribuidos de control inteligente, es permitir al sistema añadir información a los elementos de control del sistema como los agentes, componentes o comunicaciones que determinen unas restricciones de calidad en la ejecución.

2.2 Ámbito

Como se comentó en la introducción del capítulo, las políticas de gestión de las calidades de servicio proporcionan el medio por el que la aplicación y el servicio pueden negociar los parámetros de calidad de servicio. El soporte que el sistema suministra a las políticas de calidad de servicio se resume en la figura 1, donde se puede ver cómo las distintas capas que atravesarán los mensajes, van a aportar una gestión de colas que afectará a los parámetros de calidad de servicio.

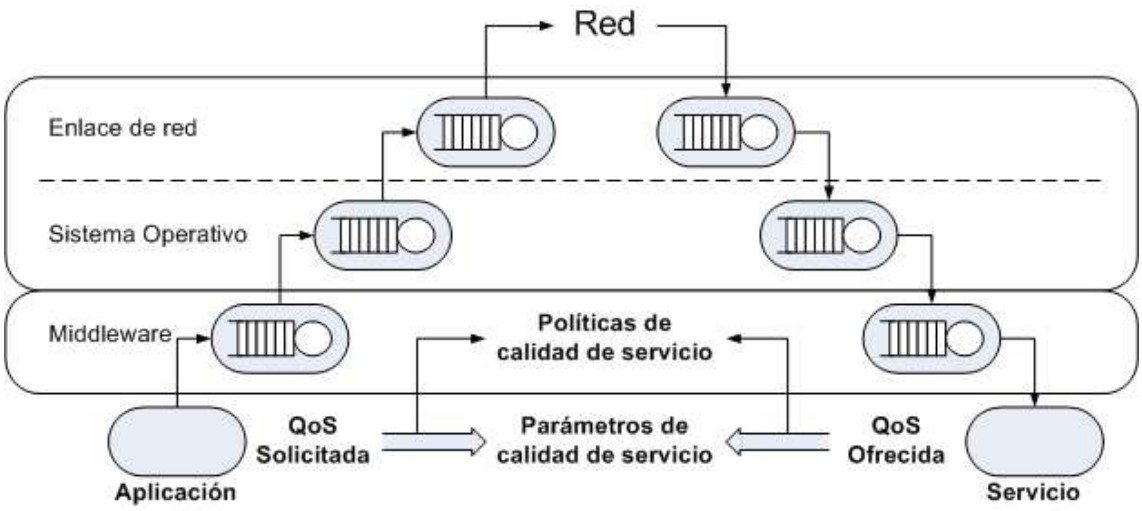

Figura 1. Relación entre parámetros y políticas de calidad de servicio.

El hecho de que las capas que atraviesa un mensaje antes de llegar a convertirse en una solicitud de servicio basen su gestión, generalmente, en colas de mensajes, hace que para revisar los parámetros más habituales empleados para definir las políticas de calidad de servicio que proporciona el middleware, se deba comenzar por los empleados por las teorías de colas.

3 Parámetros

3.1 Parámetros en las colas de mensajes

Los parámetros son las variables que, en una familia de elementos, sirven para identificar cada uno de ellos mediante su valor numérico. En el campo de la calidad de servicio no hay un consenso sobre el conjunto de parámetros que pueden emplearse para determinar si un servicio cumple con unos requisitos de calidad definidos. Por ello parece conveniente revisar distintos autores de distintos ámbitos para delimitar los más habituales y posteriormente definirlos. Cabe destacar que en la RFC 2330 [Paxson et al., 1998] se da una visión en la que se diferencia la composición de parámetros, entre una composición espacial de métricas empíricas y una composición temporal. Esta visión es interesante, como se verá posteriormente, ya que dependiendo del parámetro de calidad de servicio del que se trate, se tiene una visión temporal y espacial.

La aproximación más básica es la que se da desde la visión de la calidad de servicio en la teoría de colas. La bibliografía es amplia, aunque una visión orientada al análisis de sistemas se da en [Stuck and Arthurs, 1984] donde los indicadores de las colas se emplean para determinar el valor del rendimiento, considerándose éste último como el único parámetro relevante. Los indicadores básicos de las colas de mensajes son importantes ya que son históricamente los primeros parámetros de rendimiento que aparecen y son el origen de los indicadores que se pueden emplear para obtener los parámetros más complejos. A lo anterior hay que añadir, que en los sistemas de comunicaciones y de control, desde los basados en el paso de mensajes hasta los basados en gestión de colas distribuidas, las colas de mensajes son la estructura base de funcionamiento. Por ello, conocer qué indicadores se tienen y cómo se definen es importante para conocer cómo obtener el resto de parámetros.

Para una cola de mensajes de un componente, se puede comprobar que es asimilable el modelo de colas M/M/1, para el cual hay una serie de valores que se obtienen directamente del propio funcionamiento de la cola y que sirven de base para la obtención de los parámetros de calidad de servicio y posteriormente para la gestión de la calidad de servicio del sistema por medio de las políticas de calidad de servicio. A continuación se describen brevemente cada una de ellas.

- Tiempo entre llegadas (T). Tiempo entre un mensaje y el siguiente. Es mucho más significativo el promedio de tiempos de llegadas entre mensajes (E[T]).
- Tasa media de llegadas (λ). Representa el número esperado de llegadas a la cola de mensajes por unidad de tiempo. Se calcula como la inversa del promedio de tiempo entre llegadas.

Ecuación 1. Obtención de la tasa media de llegadas en las colas de mensajes.

$$\lambda = 1 / E[T] \quad (1)$$

- Tiempo de servicio (S). Tiempo que se tarda en atender completamente una petición. Es mucho más significativo el promedio de tiempos de servicio (E[s]).
- Tasa media de servicio (μ). Representa el número esperado de mensajes servidos por el componente por unidad de tiempo. Se calcula como la inversa del promedio de tiempos de servicio.

Ecuación 2. Obtención de la tasa media de servicio en las colas de mensajes.

$$\mu = 1 / E[S] \quad (2)$$

- Tiempo de respuesta (R). Define el tiempo de respuesta de un componente en procesar un mensaje. Es mucho más significativo el promedio de tiempo de respuestas en un intervalo de tiempos (E[T]).
- Tiempo de espera en cola (W). Especifica el tiempo que tarda un mensaje en estar en la cola esperando ser procesado. Se emplea más a menudo el promedio de tiempo de espera en cola de los mensajes en un intervalo de tiempo (E[W]).
- Número de mensajes en el nodo (N). Especifica el número de mensajes que se encuentran en un momento concreto en el componente, tanto en cola como sirviéndose. Es más relevante el valor del promedio temporal (E[N])
- Número de mensajes en cola (Nq). Define el número de mensajes que se encuentran esperando en la cola en un instante de tiempo concreto. Se emplea más el valor del promedio en un intervalo temporal (E[Nq]).
- Número de mensajes en servicio (Ns). Expresa el número de mensajes que se están procesando en el componente en un instante de tiempo. Generalmente en las colas de mensajes más sencillas es uno.
- Carga en el componente (ρ). En las colas básicas también se habla de éste indicador como tráfico. Representa la cantidad de demanda sobre el uso del componente. Se obtiene por medio de la ecuación . Expresada como porcentaje representa la ocupación.

Ecuación 3. Obtención de carga en un componente con una cola de mensajes.

$$\rho = \lambda / \mu \quad (3)$$

- Utilización del nodo (U). La utilización es el parámetro que indica el aprovechamiento del componente. En el caso de las colas sencillas M/M/1, la utilización es igual a la carga. Si la utilización se expresa en porcentaje, se habla de ocupación.
- Productividad (λe). En los sistemas de colas, la productividad se refiere a la eficiencia en el uso de la cola. En las colas sencillas la productividad tiene una relación directa con la utilización. Formalmente, lo cual es interesante para colas más complejas y sistemas de colas, la productividad se calcula a partir de las probabilidades de ocupaciones. La fórmula para las colas sencillas se puede ver en la ecuación .

Ecuación 4. Obtención de la productividad partiendo de la probabilidad de ocupación.

$$\lambda e = (1 - p_0)\ \mu \quad (4)$$

En la fórmula, p_0 es la probabilidad de encontrar vacía la cola del componente. En el caso de las colas M/M/1 ésta probabilidad se obtiene por medio de la ecuación .

Ecuación 5. Obtención de la probabilidad de i trabajos en una cola de mensajes.

$$p_i = (1 - \rho) \cdot \rho^i \quad (5)$$

En las colas de mensajes, los parámetros anteriores definen una serie de ecuaciones interesantes como la condición de estabilidad ($\lambda > \mu$) o las leyes de Little que relacionan los promedios de números de mensajes con los promedios temporales de las colas. Para obtener más información se recomienda consultar la bibliografía, hay títulos como [Averill and Kelton, 2000] donde se detallan los parámetros para sistemas más complejos.

3.2 Parámetros en los sistemas de comunicaciones

Extendiendo los conceptos de la teoría de colas a los sistemas basados en servicios, se entiende como servicio lo que en teoría de colas se entiende como sistema. En este caso el servicio mantiene las colas y los recursos a los que se acceder. En función de estos conceptos en [Stuck and Arthurs, 1984] aparecen los siguientes indicadores y parámetros.

- Tiempos de espera. Basados en los retardos que sufren los clientes al acceder a un servicio. Normalmente no se considera tiempo de espera el tiempo que el cliente disfruta del servicio.
- Carga. Se considera en función de la intensidad del tráfico de acceso a las colas para acceder a un servicio.
- Utilización. La utilización se considera como el uso que se hace de los recursos a los que da acceso el servicio, sin tener en cuenta las colas de espera.
- Capacidad. Se entiende desde el punto de vista del servicio como la cantidad de peticiones que es capaz de mantener, teniendo en cuenta las colas de espera y los recursos a los que se accede desde las colas.
- Rendimiento o producción. Se considera como el único parámetro de calidad de servicio propiamente dicho. Desde el punto de vista de las colas es la tasa de peticiones y la tasa de servicio, aunque en ocasiones se considera como la ratio de solicitudes de servicio atendidas en un intervalo de tiempo.

Cuando se aplica la teoría de colas a la medición del rendimiento de los computadores, aparecen algunos parámetros más. En [Jain, 1991] se hace una exposición de nuevos parámetros adaptados a los sistemas informáticos, considerándolos éstos como recursos.

Al igual que la obtención de parámetros en la teoría de colas, en los sistemas informáticos, los parámetros basan su cálculo en función de los tiempos que comportan las solicitudes que se realizan. En la figura 2, se puede ver los tiempos más habituales empleados como indicadores a partir de los cuales obtener parámetros de calidad de servicio.

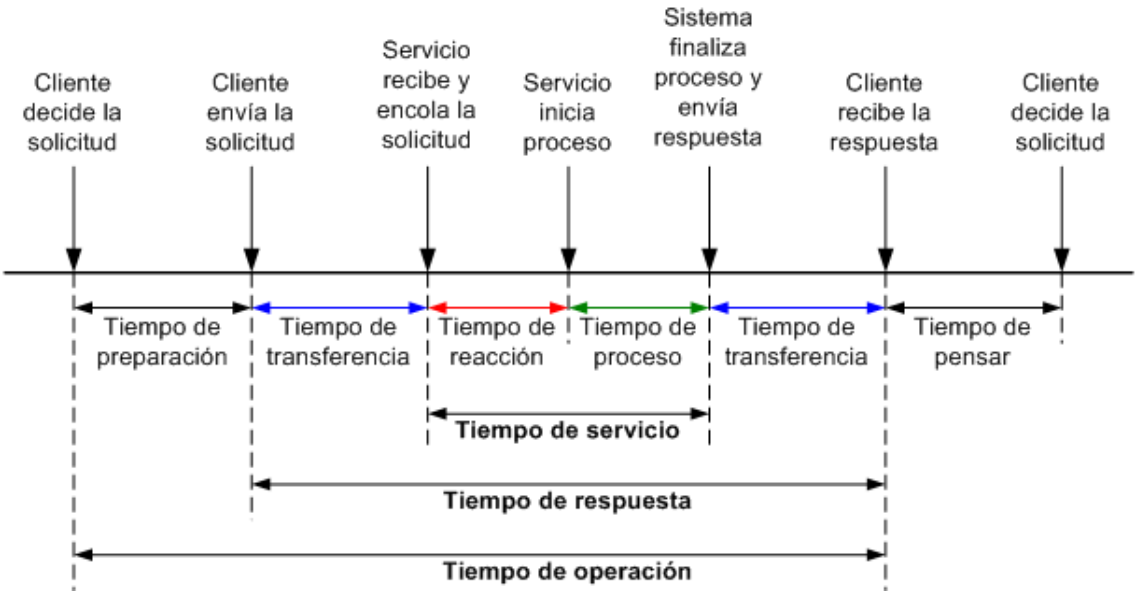

Figura 2. Análisis de los tiempos en el uso de un servicio.

Los parámetros temporales pueden detallarse más si se tiene en cuenta que la información no tiene un tamaño tan pequeño como para no tenerlo en consideración, es decir, no se transmiten y procesan bits, sino cadenas considerables. Para simplificar los cálculos no es habitual tener en cuenta estos tiempos, sino usar la aproximación más estricta. Por ejemplo, cuando se desea enviar un mensaje de una cierta longitud, el tiempo de transmisión puede considerarse desde que se comienza a enviar el primer bit, hasta que este primer bit llega al destino, o bien desde que se envía el primer bit hasta que el último que se envió llega a su destino. Como el segundo caso requiere un tiempo mayor suele ser éste último el tiempo que se considera como tiempo de transmisión. La misma filosofía se emplea en el cálculo del resto de los tiempos. Los tiempos que se consideran relevantes en [Jain, 1991] son los siguientes.

- Tiempo de respuesta. Se considera el tiempo que pasa desde que comienza a enviarse la solicitud hasta que se recibe la respuesta.
- Tiempo de operación. Tiempo de respuesta, pero incluyendo todo el proceso que interviene en la comunicación, como la preparación de los datos, operaciones y acciones similares.
- Tiempo de reacción. Tiempo que el servicio tarda en comenzar a actuar ante la solicitud. Este tiempo se debe generalmente al almacenamiento en las colas de mensajes, traducción de mensajes y aspectos similares.
- Tiempo de servicio. Tiempo total que tarda el servicio en atender una solicitud, incluyendo el tiempo en el que la solicitud está en espera de ser atendida.
- Tiempo de ocio. Tiempo en el que el servicio no es utilizado.
- Utilización. Fracción de tiempo durante la cual el servicio se encuentra ocupado sirviendo una solicitud. La utilización puede considerarse incluyendo los tiempos de espera en cola, o simplemente los tiempos en los que se está procesando el servicio.
- Disponibilidad. Fracción de tiempo durante la que el servicio está disponible. Generalmente el promedio del tiempo durante el cual el servicio no está

disponible, se conoce como “downtime” mientras que el promedio del tiempo durante el que el servicio está disponible, también se conoce, como “uptime”. A menudo, el tiempo de “uptime” se conoce como MTTF (Mean Time to Failure” o fiabilidad.

- Factor de holgura. (Stretch factor). Ratio del tiempo de respuesta en relación a la carga del sistema.

A partir de los tiempos expuestos anteriormente, puede calcularse otra serie de parámetros. Esta serie de parámetros suelen relacionar el número de servicios con parámetros temporales, estos son los siguientes.

- Rendimiento o productividad. (más conocido por el nombre en inglés thoughput). Se considera una medida de producción basada en una unidad básica, generalmente prorrateada en un intervalo de tiempo. Algunas de las más empleadas son: millones de instrucciones por segundo (MIPS), millones de operaciones en coma flotante por segundo (MFLOPS), paquetes por segundo (pps), bits por segundo (bps) o transacciones por segundo (TPS).
- Capacidad nominal. Máximo throughput.
- Capacidad de uso. Máximo throughput alcanzable sin exceder un determinado tiempo de respuesta. En ocasiones se considera también que la capacidad de uso es la carga del sistema.
- Ancho de banda. Capacidad nominal de una red.
- Eficiencia. Ratio entre el máximo throughput alcanzable (capacidad de uso) y la capacidad nominal.

Con el conjunto de parámetros vistos anteriormente se puede caracterizar la calidad de servicio. A medida que ha ido evolucionando el concepto de calidad de servicio se han ido añadiendo algunos parámetros más. En [Coulouris et al., 2001] aparecen, además del ancho de banda, ya mencionado anteriormente, los siguientes.

- Latencia. Se llama latencia al tiempo de respuesta del sistema, pero desde el punto de vista del cliente que solicita un servicio. El tiempo de respuesta es la latencia desde el punto de vista del servicio.
- Variación de la latencia. Más conocido por su denominación en inglés (jitter). Empleado para los sistemas sensibles a los cambios de la calidad de servicio a lo largo de una sesión de transmisión, como los sistemas multimedia.
- Tasa de pérdidas. Tasa de envíos erróneos en función de los envíos totales.

Los valores de los parámetros pueden ser de tres tipos [Jain, 1991]

- HB (Higher is better). Es el caso en el que para que se ofrezca calidad, cuanto más alto el valor mejor. Un ejemplo es el ancho de banda.
- LB (Lower is better). Caso opuesto al anterior, en el que la calidad se entiende como un valor bajo del parámetro. Un ejemplo es la tasa de errores.
- NB (Nominal is best). Caso intermedio en el que se debe mantener cercano a un valor, o bien dentro de un margen acotado superiormente e inferiormente.

A partir de las definiciones anteriores, que muestran la evolución temporal en la caracterización de los parámetros, y el tipo de valor que se le pide que tenga para

considerar un servicio de calidad, se pueden crear políticas de calidad de servicio. Las características comunes pueden verse en la figura 3.

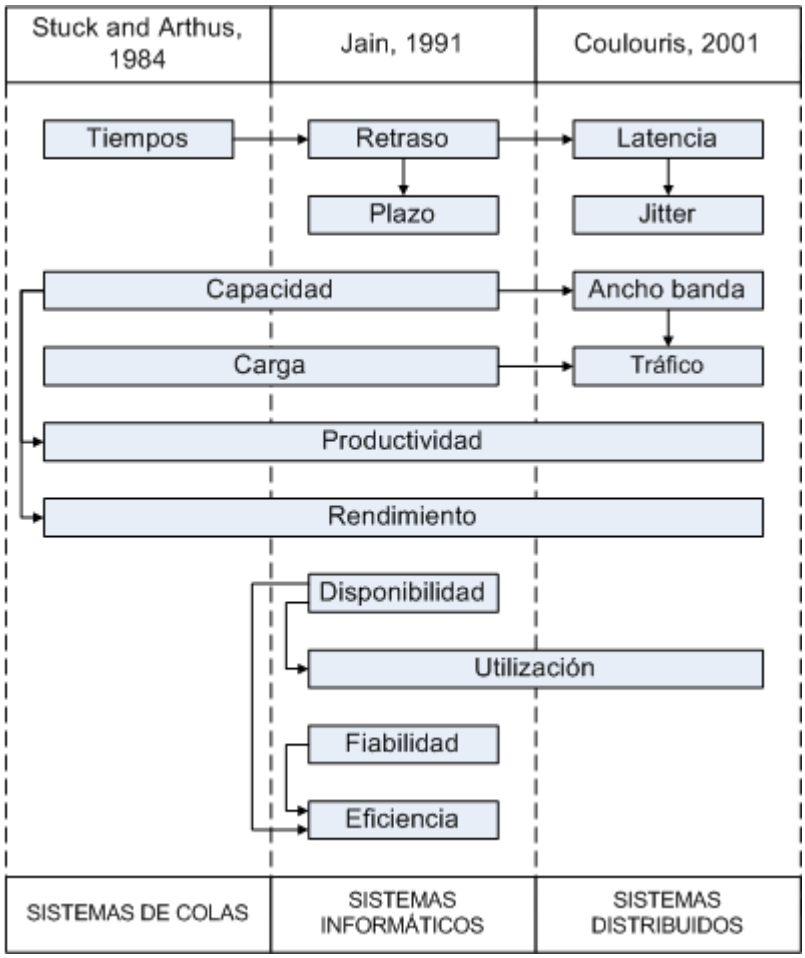

Figura 3. Parámetros de calidad de servicio en función del ámbito en el que se aplican.

En la figura 3, se observa la evolución de los parámetros que se han considerado relevantes en función del tipo de sistema sobre el que se aplican. Cabe destacar que en los sistemas distribuidos se habla de latencia en lugar de retrasos y de tráfico en lugar de la carga. Esto no es incongruente, ya que en los sistemas distribuidos, además de los parámetros de calidad de servicio propios de los nodos (sistemas informáticos y de colas a todos los efectos) se debe añadir la medición de los parámetros propios de la red. También cabe destacar cómo los tiempos, las cargas y consecuentemente el rendimiento son comunes a los tres ámbitos mostrados. A continuación se da una definición más detallada y común de las principales características basada en las definiciones anteriores.

3.2.1 Tiempo de espera

En inglés se conoce como “Waiting time”. El tiempo de espera es uno de los parámetros basados en los sistemas de colas. Se calcula siempre en función de la cualidad sobre la que se trate, en el caso del procesamiento se suele referir al tiempo que se tarda entre que se solicita un procesamiento y se obtiene el resultado. En el caso de las comunicaciones suele hablarse del tiempo que tarda la respuesta en llegar ante el envío de un mensaje. A un nivel más alto, en los sistemas de control se puede tener diferentes visiones dependiendo de lo que se deba “esperar”, en el caso del resultado de una acción

se puede hablar del tiempo de espera como el tiempo transcurrido entre que se recibe un cambio en la entrada de información (sensores) se procesa la entrada, se toma una decisión y se actúa sobre la salida de información (actuadores)

Como parámetro se considera una magnitud de dimensión temporal que se mide en las unidades temporales correspondientes al sistema. En una visión general suele expresarse como promedio de los tiempos de respuesta del sistema, lo que relaciona al tiempo de espera y el tiempo de respuesta del sistema, considerándose en algunos casos el mismo parámetro.

3.2.2 Retardo

En inglés se conoce como "*delay"* y en ocasiones "*hold time*". Tiempo que transcurre entre el instante de tiempo esperado y el instante real en el que sucede un acontecimiento. El retraso tiene muchos posibles significados, pero cuando se habla de QoS, se suele relacionar con alguna de las características sobre las que se trabaja. Por ejemplo, retraso de procesamiento es el tiempo que transcurre entre que un componente recibe un mensaje y el mensaje es procesado y se realiza la acción correspondiente.

Desde el punto de vista de las comunicaciones se entiende como retraso el tiempo que se tarda desde que un dispositivo recibe una trama hasta que la trama es retransmitida hacia el puerto correspondiente. Se suele hablar también de "retraso de serialización", como el tiempo que se tarda en transmitir un paquete o trama. Y retraso de extremo a extremo (*end to end*) como el retraso total que los paquetes experimentan desde la fuente al destino, es decir desde que inician su trayecto hasta que llegan al destino final.

3.2.3 Plazo temporal

En inglés *deadline*. Aunque no es considerado como factor de calidad de servicio propiamente dicho, es el cumplimiento del "deadline", lo que debe considerarse como un factor de calidad de servicio.

Para comprobar si un sistema cumple los plazos temporales, se deben cumplir los retrasos correspondientes de transmisión y de procesamiento.

3.2.4 Latencia

(*latency*). Tiempo que transcurre entre que se da una acción y se produce la respuesta. En [Jain, 1991] se le llama "Tiempo de Respuesta". En el caso de las comunicaciones se suele calcular como el tiempo transcurrido mientras una unidad de datos entra en el componente correspondiente y lo abandona. En ocasiones se trata como sinónimo de "retraso", aunque formalmente sean distintos conceptos ya que el retraso implica una desviación sobre una expectativa de tiempo, mientras que la latencia es un valor proporcionado por el sistema.

3.2.5 Inestabilidad

Inestabilidad (*jitter*). Se define como la ausencia de cambios o de variaciones en función de una referencia. En el ámbito de las redes se habla de un cambio o variación de la latencia. También se considera la variación del retraso. Se considera un mal término y se recomienda no emplearlo ya que dependiendo de qué entorno se trabaje se define de formas diferentes. En [Demichelis and Chimento, 2002] se habla de abandonar el uso y emplear PDV (Packet Delay Variation) como parámetro de calidad de servicio.

3.2.6 Capacidad

Capacidad (*capacity*). La capacidad suele entenderse como un máximo alcanzable en un sistema. En [Jain, 1991] se diferencia la capacidad nominal (máximo rendimiento) y la capacidad de uso (máximo rendimiento para un tiempo de respuesta dado). Esta diferenciación es importante, ya que la capacidad de uso es la medida más realista sobre la capacidad que puede proporcionar el sistema que cumpla los requisitos temporales que se le exigen al servicio.

3.2.7 Ancho de banda

Ancho de banda (*bandwidth*). Se define en sí misma como la capacidad de producción o rendimiento prorrateada de una red, un medio de comunicación o protocolo de comunicaciones. En el caso de la calidad de servicio, suele hablarse más del reparto del ancho de banda, ya que como calidad de servicio no se tiene la capacidad de acceder al medio del que depende el ancho de banda. En ocasiones se llama ancho de banda a la velocidad o al rendimiento de un medio de comunicación. Como término ambiguo no suele recomendarse su utilización, sino los parámetros que puedan sustituirle.

3.2.8 Carga

Carga (*load*). Normalmente se refiere a la cantidad de peticiones en el caso de las colas, de procesos en el caso de los sistemas o de mensajes en el caso de las comunicaciones. Generalmente es una medida de la cantidad de trabajo que tiene un sistema.

3.2.9 Tráfico

Tráfico (*trafic*). Se entiende como la carga pero desde un punto de vista dinámico, es decir la carga a lo largo de un tiempo, suele aplicarse a las redes de comunicaciones, con la unidad de medida de la información (bytes, paquetes, mensajes y similares) a lo largo de un intervalo de tiempo definido.

3.2.10 Utilización

Utilización (*utilization*). Se define como la tasa de empleo o uso de algo. Es una medición más exacta que la productividad. Suele calcularse como el porcentaje de tiempo durante el cual un servicio (o recurso) esta ocupado. En el ámbito de las comunicaciones se emplea para localizar zonas de congestión y cuellos de botella a partir de valores elevados de utilización de componentes.

3.2.11 Productividad

Productividad (*throughput*). La productividad es uno de los parámetros más empleados y con más posibles interpretaciones. Se define como la capacidad o el grado de producción por unidad de trabajo y suele calcularse como la relación entre lo producido y los medios empleados o si se desea obtener en función del tiempo se calcula como la cantidad de trabajo realizado en un periodo de tiempo. En el ámbito de las comunicaciones, suele hablarse de la cantidad de información que una red es capaz de entregar durante un intervalo de tiempo.

3.2.12 Rendimiento

Rendimiento (*performance*). Formalmente el rendimiento se define como la "proporción entre el producto o el resultado obtenido y los medios utilizados" y la

productividad se define como la "relación entre lo producido y los medios empleados" por lo que pueden considerarse similares. En ocasiones se expresa el rendimiento como un porcentaje de la productividad, lo que hace que pueda localizarse en ocasiones de forma diferenciada a la productividad.

3.2.13 Disponibilidad

Disponibilidad (*availability*). La disponibilidad se define como la cualidad o condición de poder disponer libremente de un servicio (o recurso) o bien que dicho servicio (o recurso) está listo para usarse o utilizarse. Se suele calcular como el tiempo mínimo que se asegura que el servicio estará en funcionamiento, en ocasiones se calcula como el tanto por cien del tiempo en el que se puede obtener el servicio si éste se solicita.

3.2.14 Fiabilidad

Fiabilidad (*reliability*). Se define como cualidad de que un servicio (o recurso) ofrezca seguridad o buenos resultados. Generalmente se calcula como la probabilidad del buen funcionamiento de algo. Para el caso de las comunicaciones, se suele concretar al campo sobre el que se aplica. Por ejemplo, se habla de "tasa de pérdida de paquetes" para referirse a la tasa de paquetes que se pierden o se descartan, por llegar alterados o por tener un retraso excesivo, en el caso de las aplicaciones de tiempo real. Formalmente se proporciona como la probabilidad del buen funcionamiento de algo.

3.2.15 Eficiencia

Eficiencia (*efficiency*). Es un término muy similar al de rendimiento, y en muchas ocasiones suelen utilizarse indistintamente. En [Jain, 1991] se define como la ratio entre el máximo rendimiento obtenible (capacidad de uso) y la capacidad del sistema (capacidad nominal). La eficiencia se refiere especialmente a la optimización de los recursos empleados, y generalmente se diferencia del rendimiento en que éste último se expresa en función de una unidad, generalmente temporal, mientras que la eficiencia se proporciona como una tasa.

3.2.16 Otros parámetros

Además de los parámetros expuestos anteriormente, existe una gran cantidad de indicadores empleados para determinar la calidad de un servicio concreto. En estos casos los parámetros suelen adaptarse a las características específicas del servicio sobre el que se desea una calidad concreta, por ejemplo en redes se suele hablar de tasas de errores, tasas de paquetes perdidos y parámetros similares que sólo se aplican en ese ámbito.

4 Extensión de parámetros básicos a parámetros avanzados

4.1 Relación entre parámetros

A partir de los indicadores básicos de las colas: tiempos y número de mensajes, es posible obtener la mayoría de los parámetros que la bibliografía entiende como los habituales en la calidad de servicio de las comunicaciones. En la figura 4, se puede ver cómo a partir de los indicadores básicos, fácilmente obtenibles por las colas de mensajes, se pueden obtener una serie de indicadores más complejos.

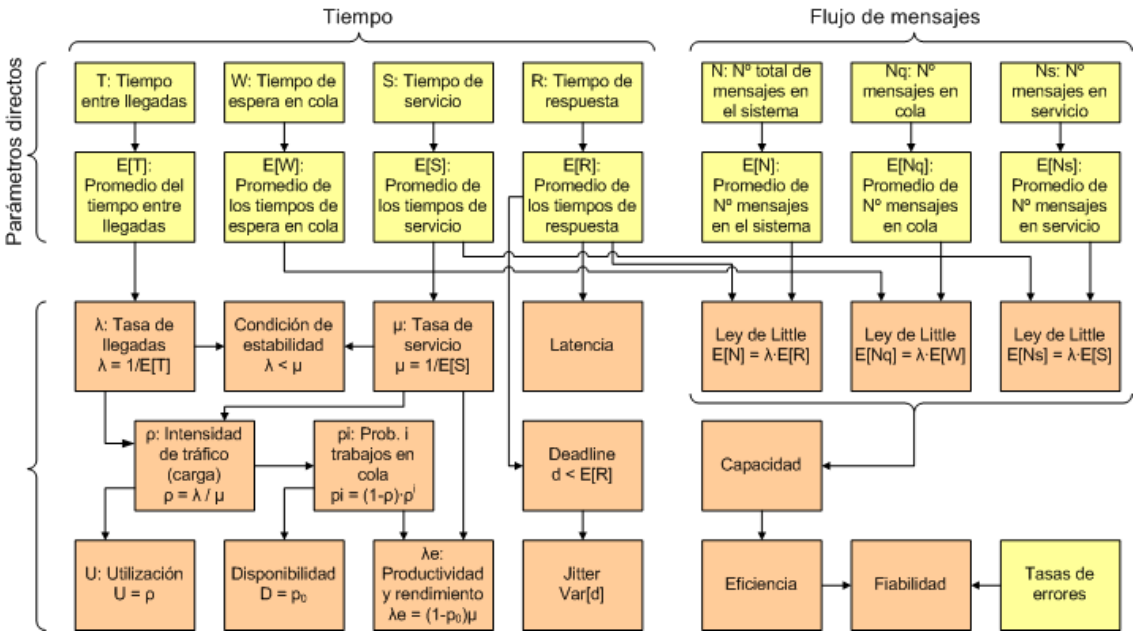

Figura 4. Parámetros obtenidos a partir de los indicadores de las colas de mensajes.

Las variables de las colas de mensajes que se pueden obtener directamente se agrupan en dos conjuntos, las variables relacionadas con los valores temporales y las relacionadas con el número de mensajes en la cola. Estos dos agrupamientos están relacionados directamente con los aspectos de gestión temporal y de control de flujo ampliamente utilizados para definir la calidad de servicio. El hecho de que la mayor parte de los parámetros de calidad de servicio pueda obtenerse por medio de los parámetros de rendimiento de las colas de mensajes demuestra la relación existente y el hecho de que el interés hacia la calidad de servicio surja también cuando los sistemas de comunicaciones pasan a emplear las colas de mensajes en todos sus componentes.

4.1.1 Parámetros básicos

Estos parámetros son los que en la ilustración se consideran directos, comprenden valores que se obtienen directamente de las colas de mensajes y elementos de cómputo de los componentes del sistema.

1. Promedio de tiempo entre llegadas. Se obtiene directamente a partir de la monitorización de los tiempos entre llegadas de mensajes. E[T].
2. Promedio del tiempo de espera en cola. Se obtiene monitorizando el instante de tiempo en que un mensaje se encola a la espera de ser procesado y el instante de tiempo en que el mensaje deja la cola para pasar a ser procesado, E[W].
3. Promedio de tiempo de servicio. Promedio de tiempo que se tarda en procesar un mensaje (una vez extraído de la cola). E[S].
4. Promedio de tiempo de respuesta. Promedio de tiempo desde que un mensaje llega a la cola hasta que se termina su procesamiento. E[R].
5. Promedio del número de mensajes en el sistema. Promedio de mensajes que están en el sistema (en cola y en proceso). E[N]
6. Promedio de mensajes en cola. Promedio de los mensajes que permanecen en cola. E[Nq]

7. Promedio de mensajes en servicio. Promedio de los mensajes que permanecen en servicio (procesándose) en el sistema. E[Ns]

Estos parámetros son sencillos de calcular y se obtienen directamente del código en los hilos de gestión de colas de mensajes de los componentes.

4.1.2 Parámetros derivados

La teoría de colas permite obtener una serie de parámetros básicos a partir de los parámetros anteriores, que no son más que los indicadores de características básicas de las colas. En concreto los parámetros son los siguientes

1. Tasa de llegadas
2. Condición de estabilidad
3. Tasa de servicio
4. Latencia
5. Ley de Little (carga del sistema y tiempo de respuesta)
6. Ley de Little (carga de colas, tiempo de espera)
7. Ley de Little (carga de proceso)
8. Intensidad de tráfico, carga
9. Probabilidad de carga
10. Deadline
11. Capacidad
12. Utilización
13. Disponibilidad
14. Productividad
15. Jitter
16. Eficiencia
17. Fiabilidad

4.2 Calidad de control y calidad de servicio

Este apartado trata de los puntos 2, 3 y 4 de las posibles ampliaciones en lo que a parámetros de QoS se refiere.

Existen diversas aproximaciones sobre parámetros que proporcionan la calidad de un sistema en algún aspecto. Por ejemplo en [Gabel and Litz, 2003] se propone como forma para el cálculo de la QoC (Quality of Control) a partir de una fórmula en la que aparecen constantes de ajuste y funciones generales sobre el comportamiento final de los componentes.

$$QoC = \frac{1}{(1-\lambda)\cdot IAE + \lambda \cdot ITAE} \cdot 1000 \qquad IAE = \int_0^{t_{END}} |y(t) - r(t)| dt \quad \text{and} \quad ITAE = \int_0^{t_{END}} |y(t) - r(t)| t dt .$$

Ilustración 1. Calidad de Control descrita en [Gabel and Litz, 2003]

En [Soucek and Sauter, 2004] se hace la aproximación de obtener la calidad de control relacionada con la QoS. En concreto se tratan los siguientes parámetros.

End to end delay Retraso total de un paquete que atraviesa una ruta específica. Es un parámetro calculado a partir del acumulado de los retrasos de los componentes que se atraviesan.

$$D_k = \sum_{e \in E(\mathrm{P})} d_e + \sum_{n \in V'(\mathrm{P})} s_{k,n}.$$

La calidad de éste parámetro se obtiene a partir del cumplimiento de unas condiciones definidas.

$$\mathrm{P}(D_i \leq D_{\max}) \geq Z_{\min}.$$

Delay Jitter Calculado con la parte estocástica del "end to end delay".

$$D_k = D_{\min} + J_k$$

La calidad del parámetro se obtiene a partir del cumplimiento de las siguientes condiciones.

$$\mathrm{P}(J_i <= j_{\max}) \geq U_{\min}.$$

Throughput Se define como la capacidad de transmisión en un determinado camino. En el caso estudiado se propone como fórmula.

$$\Theta = \frac{A(t+\Delta t) - A(t)}{\Delta t}.$$

Donde A(t) se define como la cantidad de datos transferidos hasta el instante t. al igual que el resto de parámetros, se consideran de calidad en las siguientes condiciones.

$$\mathrm{P}(\Theta \geq \Theta_{\min}) \geq \zeta_{\min}.$$

Loss rate (reliability) Se refiere a la clásica tasa de pérdidas de paquetes, generalmente debida a congestiones. Suele llamarse como confiabilidad. El parámetro se define directamente por el cumplimiento del umbral de un valor.

$$\mathrm{P}(packet\ received) \geq W_{\min}.$$

Packet ordering Se considera como parámetro la llegada en el orden adecuado.

$$\langle p_k \rangle = \langle p'_k \rangle \qquad \forall k.$$

Cabe destacar que en [Soucek and Sauter, 2004], se consideran como parámetros de QoS el jitter inducido por la red y por el protocolo. Además se crea un vector de calidad de servicios.

$$\mathbf{q}^T = (D_{\min}, J_k, 1-w).$$

Si se considera el control como un servicio, y se tiene en cuenta la definición de QoS donde se habla de efecto colectivo, parece conveniente averiguar la posibilidad de obtener la calidad del control a partir de los parámetros básicos y de los componentes específicos.

4.3 Parámetros de calidad de servicio de DCPS

El modelo DCPS del estándar DDS, proporciona unas funciones que implementan el comportamiento del sistema, es decir "usan" la calidad de servicio para definir un comportamiento del sistema, a eso le llaman "política de calidad de servicio". Las políticas de calidad de servicio de DCPS del estándar DDS pueden "insinuar" algunos parámetros de calidad de servicio. A continuación se revisan éstos posibles parámetros.

- **Durabilidad**. Número de muestras de mensajes que se deben mantener en cada nodo. Permite mantener un historial de mensajes
- **Deadline**. Plazo temporal que se debe mantener para que un mensaje sea válido. Define el retardo máximo que puede tener el mensaje.
- **Latencia**. La latencia (en DCPS) se define como el retraso admisible, pero a nivel global (no sólo como el deadline), se podría decir que es un deadline a nivel global.
- **Vivacidad** (liveliness). Capacidad para comprobar qué elementos del sistema están "vivos". Es un parámetro que permite localizar componentes que no están accesibles (ya que no envían mensajes de liveliness.
- **Filtro temporal**. (periodo). Separación mínima que deben tener los mensajes. Es el parámetro complementario al deadline (separación temporal máxima que podemos permitir entre mensajes), entre los dos parámetros se puede establecer algo similar al "Jitter".
- **Lifespan**. Vida útil. Especifica el tiempo que un dato puede ser considerado válido. Es muy interesante ya que permite determinar qué tiempo podemos permitir que un dato esté activo.
- **Presentation**. Se refiere a dos características.
    - **Coherent access**. El acceso coherente se refiere al ámbito en que afectan los datos.
    - **Orderer access**. Se refiere a la importancia que se le da a los accesos ordenador.
- **Ownership**. Propiedad o posesión. Se trata de especificar qué componentes pueden escribir datos en otros componentes.
- **Ownership strength**. Realmente debería ser prioridad, pues se refiere a la prioridad en el acceso a componentes que tienen otros componentes.
- **Partition**. Se refiere a áreas de comunicación aisladas frente a otras, puede tener interés en cuanto a aislar unos componentes de otros.
- **Reliability**. Fiabilidad. En DCPS éste concepto se refiere a la propagación, es decir a si se permite llegar datos más recientes antes que otros que no han sido recibidos.
- **Transport priority**. Prioridad de los mensajes.
- **Destination order**. Se refiere al orden en que se deben procesar los mensajes, si en el orden de salida o en el orden de llegada (que ha podido ser variada en la ruta)
- **History**. Número de muestras que se mantiene en cada componente.

- **Resource limits**. Se encarga de conocer los límites de recursos que se deben gestionar en cada componente.

4.4 Parámetros de calidad de servicio en sistemas distribuidos

Para adaptar los parámetros de calidad de servicio a los sistemas de los que se tratan es necesario conocer las características de los sistemas en los que se pretender aplicar. En [Poza, et al., 2009] se realiza una revisión de los sistemas distribuidos de control, donde se proponen una serie de características que éstos sistemas deben tener.

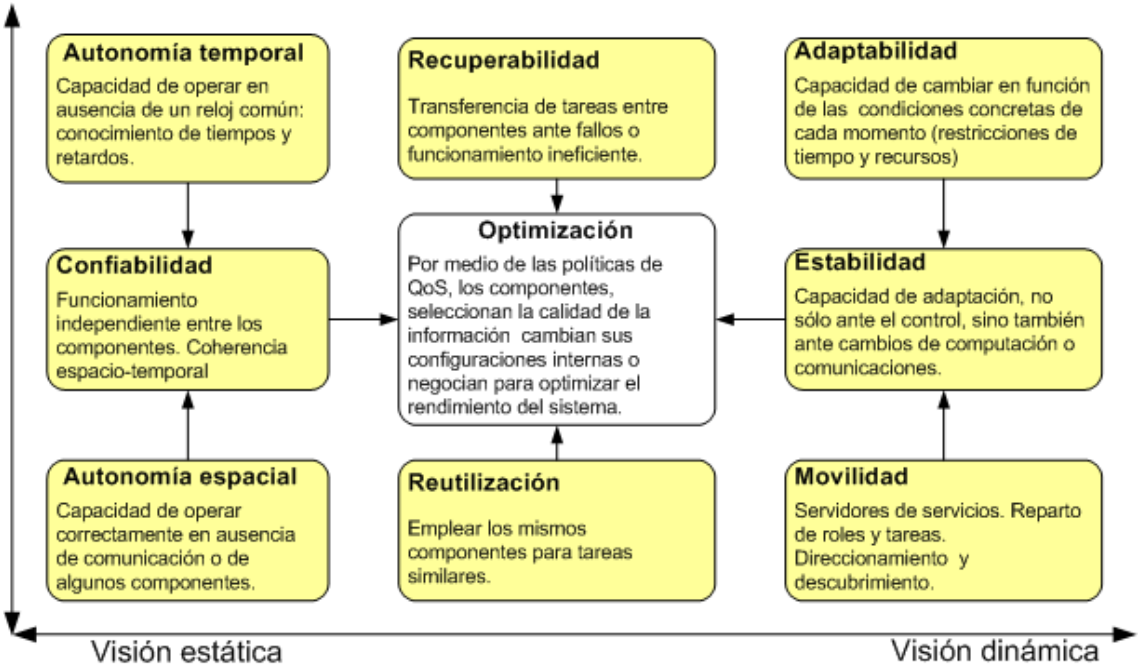

Ilustración 2. Características deseables para la optimización de un sistema.

A partir de los componentes de la **Ilustración 2**, se pueden extraer una gran cantidad de posibles parámetros que indiquen el nivel de calidad en la optimización del sistema. Algunos de éstos podrían ser:

- Componentes que den soporte a las características anteriores. Esto hace que se pueda justificar la arquitectura desarrollada.
    - Tipos de componentes.
    - Detalles de cada componente.
    - Interacciones entre los componentes
    - Evolución de los componentes (estados en los que se encuentran los componentes)
    - Evolución de las interacciones entres los componentes (protocolos de comunicación entre los componentes)
- Índices que permitan cuantificar las cualidades definidas por las características: cabe destacar que hay características fácilmente "cuantificables", como la estabilidad y otros más difícilmente cuantificables, como el direccionamiento semántico. Una primera aproximación de los índices posibles (unos más y otros menos) sería la siguiente.

- Índices de autonomía temporal. Deben indicar la capacidad que tiene un componente[1] de no depender de los aspectos temporales del sistema. Una de las formas más sencillas de evitar el uso de un reloj común, lo que aísla a los componentes de un servidor de tiempos. Otra de las formas es proporcionar a los componentes la capacidad de conocer tiempos de procesamiento y retardos de las comunicaciones. Umbrales de tiempo que se puede trabajar sin tener mensajes. Capacidad de predicción del control en ausencia de entradas (margen temporal)
- Índices de autonomía espacial[2]. Indican la capacidad de operar correctamente en ausencia de algunos componentes. La ausencia de los componentes puede producirse por diversos motivos: error en las comunicaciones, caída de temporal parcial o total de los componentes, retardos esporádicos excesivos, etc.
- Índices de confiabilidad. La confiabilidad es cuando se dan las características espacio-temporales expuestas anteriormente, es decir cuando los componentes pueden funcionar de una forma más o menos independiente unos de otros. También se considera como índice de coherencia espacio-temporal, es decir la capacidad de poder reconocer que los resultados de los componentes se corresponden de forma coherente a lo esperado.
- Índices de reutilización. La reutilización es posibilidad de emplear un mismo componente genérico para la misma tarea. La reutilización está relacionada con la movilidad (reutilización del mismo componentes en diferentes ubicaciones) y con la clonación (reutilización simultánea del mismo componente).
- Índices de adaptabilidad. La adaptabilidad es la capacidad de los componentes a cambiar sus requerimientos en función de las condiciones concretas de cada entorno. Formalmente es la capacidad de reaccionar a los cambios del entorno para realizar la misma tarea con una calidad concreta.
- Índices de estabilidad. La estabilidad está relacionada directamente con la adaptabilidad. Si a la adaptabilidad se le incluye la variable temporal, entonces se tiene la estabilidad, es decir la capacidad de reaccionar en un tiempo adecuado para que no afecte a los parámetros especificados de funcionamiento.
- Índices de optimización. Posiblemente los índices más interesantes e importantes sean los que tratan de optimizar el sistema (hacerlo, adaptable y estable) hacia unas características concretas

[1] Hay que tener en cuenta que los índices pueden ser de los componentes o del sistema. Realmente si el índice de un componentes es IC_i, el del sistema es $f(IC_1,\ldots IC_N)$, para N componentes del sistema. Esto da mucho juego, pues permite obtener valores de índices de componentes globales, en función de los valores de los componentes individuales.

[2] Realmente no es una autonomía espacial física, sino del "espacio de componentes" que intervienen o que se precisan para realizar una acción de control concreta.

  - Índices de recuperabilidad. Cuando el sistema está en funcionamiento, se requiere que se tenga la capacidad de variar en función de estímulos exteriores[3] no habituales, se podría generalizar considerando que es la capacidad de volver a funcionar ú optimizarse a partir de errores del sistema.
  - Índices de direccionamiento. También se podría hablar de índices de localización espacial, o de "espacialización" de los componentes, es decir la capacidad de clonarse (copiar en un destino) y de moverse (copiar en un destino y borrar en el origen).

A las áreas de los índices anteriores se les pueden ir asociando diversos índices. De la revisión bibliográfica realizada en la primera parte de éste documento, se pueden deducir algunos de los índices más empleados en los sistemas de control distribuido.

- Carga de trabajo
- Carga de bus
- Uso del procesador
- Tiempos de respuesta
- Cargas de las transmisiones
- Coste del sistema
- Estabilidad
- Determinismo: Capacidad de predicción, determinación de parámetros de las acciones de control con márgenes de confiabilidad correctos.
- Comportamiento del peor caso
- Rendimiento de tiempo prometido: tiempo negociado, posiblemente se trata de un tiempo de respuesta comprometido después de la negociación
- Variación de los retardos (jitter)
- Control de flujo (parámetro de QoS)
- Probabilidad de transmisión de un componente.
- Rendimiento
- Retraso
- Fairness: Imparcialidad, equidad, proporcionalidad entre los componentes.
- Índices de autonomía espacial. o Posiblemente estos índices estén ya reflejados en deadline, etc. de DCPS. Índice de capacidad de aislamiento
- Índices de autonomía temporal. Posiblemente estos índices estén ya reflejados en deadline, etc. de DCPS
  - Índice de autonomía temporal (índice de autonomía)

[3] Realmente se podría decir que la estabilidad es un concepto interno (auto-optimización a partir de la configuración interna y del funcionamiento normal) mientras que la recuperabilidad es un concepto externo (auto-optimización en función de variaciones no inherentes al sistema, estímulos externos).

- Índices de confiabilidad (posiblemente habría que obtenerlo a partir de los índices de autonomía espacial y temporal).
- Índices de recuperabilidad
- Índices de reutilización
- Índices de adaptabilidad
- Índices de estabilidad
- Índices de movilidad
- Índices de optimización

A partir de estos posibles índices se puede buscar la ubicación y posible relación con los parámetros más clásicos que se han hecho en las revisiones previas.

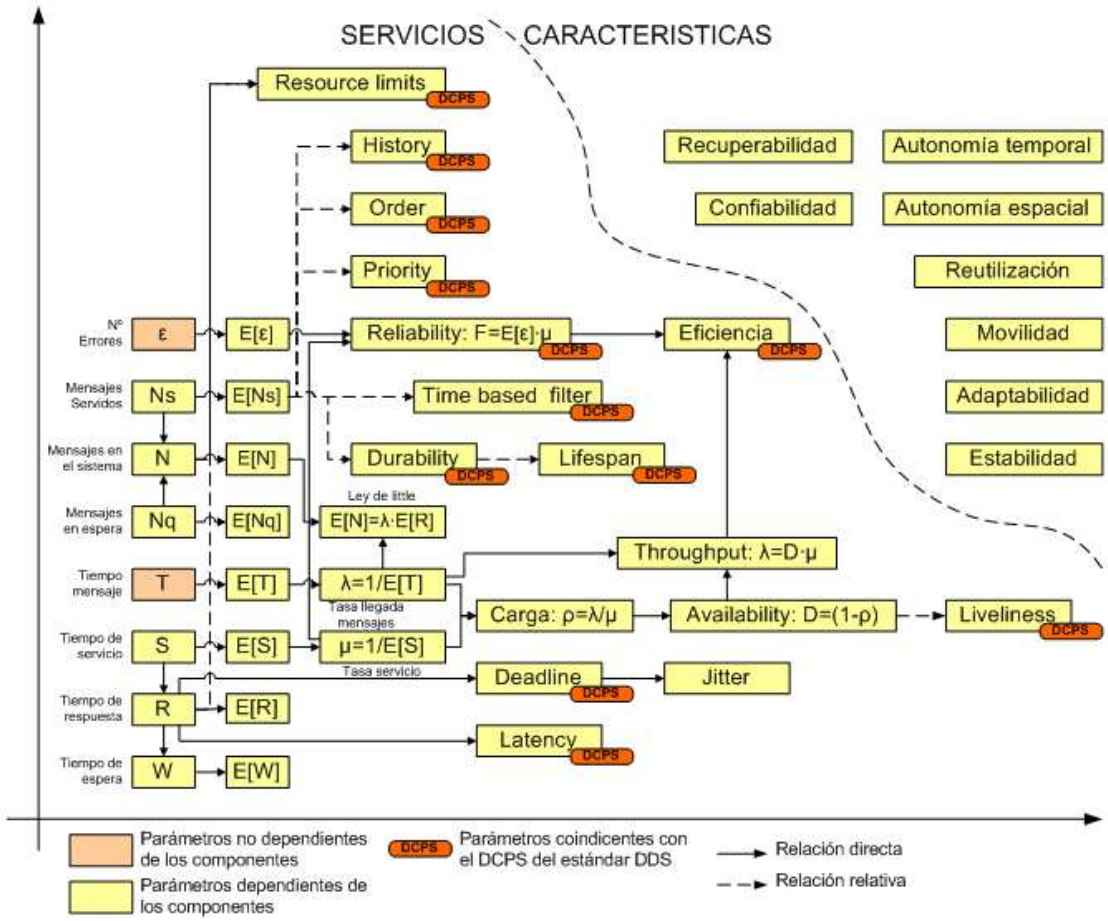

Ilustración 3. Ubicación en relación a la complejidad de algunos parámetros de calidad de servicio.

A partir de algunos de estos índices se puede empezar a buscar las fórmulas que pueden implementarlos y la experimentación de la relevancia que tienen estos índices realmente en la optimización.

5 Conclusiones

5.1 Calidad de servicio en las comunicaciones

La calidad de servicio es una herramienta de gran utilidad para la evaluación y el ajuste en el funcionamiento de un sistema de comunicaciones. Desde los parámetros más básicos de las colas de mensajes, hasta los parámetros complejos todos dan una apreciación del estado de los servicios que el sistema debe proporcionar.

Las relaciones entre parámetros de calidad de servicio permiten afrontar la obtención de parámetros complejos a partir de parámetros básicos. Es por ello por lo que la medición de los parámetros básicos podrá ser utilizada, en cierta medida, para obtener parámetros más complejos que den información de sistemas, a su vez, complejos.

5.2 Posibles líneas de investigación

La calidad de servicio puede ser considerada más una herramienta de apoyo a la evaluación y el funcionamiento de un sistema, que una línea de investigación concreta. A pesar de todo, resulta de gran interés el análisis teórico de la incidencia del uso de políticas de calidad de servicio en los sistemas.

Sin embargo, la investigación en la determinación de los índices de calidad de los sistemas es, posiblemente, uno de los campos donde se puede desarrollar ciertas líneas de investigación. **A medida que crece la complejidad de un sistema, crece la complejidad de los índices que determinan su buen funcionamiento**. Por ello, la correcta determinación de índices de calidad en función del objetivo en el que se aplican es un campo interesante de investigación.

6 Referencias

Averill and Kelton, 2000 — Averill M. Law y W. David Kelton. Simulation Modeling and Analysis. Tercera edición. McGraw-Hill, 2000.

Coulouris et al., 2001 — Coulouris, G., Dollimore, J., Kindberg, T. Sistemas Distribuidos, Conceptos y diseño. Tercera Edición. Addison Wesley. Madrid. 2001.

Crawley et al., 1998 — Crawley, E.; Nair, R; Rajagopalan, B. "RFC 2386: A Framework for QoS-based Routing in the Internet". August. 1998, pp. 1-37, XP002219363

Demichelis and Chimento, 2002 — Demichelis, C.; Chimento, P. "RFC 3393: IP Packet Delay Variation Metric for IP Performance Metrics (IPPM)". The Internet Society. November 2002.

Gabel and Litz, 2003 — Gabel O., Litz L., "QoS-adaptive Control in NCS with Variable Delays and Packet Losses – A Heuristic Approach", 43rd IEEE Conference on Decision and Control, 2004.

ITU, 1994 — ITU-T Recommendation E.800 (0894). Terms and Definitions Related to Quality of Service and Network Performance Including Dependability, 1994.

Jain, 1991 — Raj Jain. The art of Computer Systems Performance Analysis. John Wiley & Sons Inc. New york. 1991

Paxson et al., 1998 — Paxson, V., Almes, G., Mahdavi, J. and M. Mathis, "Framework for IP Performance Metrics", RFC 2330, Mayo 1998.

Poza, et al., 2009 — J.L. Poza, J.L. Posadas y J.E. Simó. Arquitecturas de control distribuido. Technical Report. Universidad Politécnica de Valencia. 2009.

Soucek and Sauter, 2004 — Soucek S., Sauter T.,"Quality of Service Concerns in IP-Based Control Systems", IEEE Transactions on Industrial Electronics, Vol. 51, No. 6, December 2004.

Stuck and Arthus, 1984 — B.W. Stuck and E. Arthurs. A Computer & Communications Network Performance Analysis Primer. Prentice Hall. 1984.

Vogel et al., 1995 — Andreas Vogel, Brigitte Kerherve, Gregor von Bochmann, Jan Gecsei. Distributed Multimedia and QoS: A Survey. Vol.2,. No. 2, 1995, pp.10-19.