Informe de Elasticidad

El siguiente es informe tiene como objetivo la realización de una práctica que sigue el modelo dado en el capítulo 8 del libro “Performance by Design: Computer Capacity Planning by Example de Daniel A. Menascé”, para construir un controlador de elasticidad que permita, a partir de una configuración, encontrar una mejor configuración para soportar la carga de operaciones de entrada al sistema.

Actividades:

1º. Construir el modelo de la aplicación. A partir del empleo de los algoritmos MVA multiclase construir la gráfica que representa (para una configuración dada) el tiempo de respuesta frente a la carga de entrada global del sistema. Realizar distintas gráficas con diferentes configuraciones y comparar los resultados.

2º Definir en base a los resultados obtenidos sobre el sistema los niveles de QoS que servirán como base para evaluar diferentes configuraciones. Básicamente se trata de definir cuál es el tiempo de respuesta objetivo, bien para cada clase de operaciones o un tiempo global medio. Para evaluar una configuración con una carga dada debéis proponer alguna función de utilidad que indique la relación del coste (número de instancias en cada capa para una configuración), el tiempo de respuesta, y la carga soportada.

3º Idear un proceso de elasticidad que se base en las medidas que es razonable extraer del sistema en funcionamiento.

Concepto de Elasticidad (según Herbst)

Es el grado por el cual un sistema es capaz de adaptarse a los cambios en la carga de trabajo que soporta mediante el abastecimiento y desabastecimiento de recursos de cómputo de forma autónoma, de tal manera que en cada punto del tiempo los recursos disponibles encajan con la demanda de recursos real tan cerca como sea posible.

Un concepto asociado es el de Escalabilidad, el cual brinda la base para que un sistema poder tener cierto grado de elasticidad. Este indica la habilidad de un sistema para reaccionar y adaptarse sin perder la calidad, también para manejar el crecimiento continuo de trabajo de manera fluida, o de una forma más sencilla estar preparado para crecer.

En el caso de la práctica nos enfocaremos en un escalado horizontal, y se lograra adicionando o eliminando componentes teniendo en cuenta en que cantidad, esto se consigue mediante un proceso de adaptación que valla indicando que decisión tomar a medida que valla cambiando.

**Elementos generales a tener en cuenta**

Ui -> utilización media del centro de servicio i

Si -> tiempo medio de servicio por operación en el centro de servicio i

Xi -> Productividad media del centro de servicio i

Vi -> Número de veces que una operación visita el centro de servicio i

X0 -> Productividad del sistema

Di -> Media de la demanda del centro de servicio i

Ri -> Media del tiempo de respuesta por operación en el centro de servicio i

R0 -> Media del tiempo de respuesta del sistema

N -> Numero de operaciones en el sistema

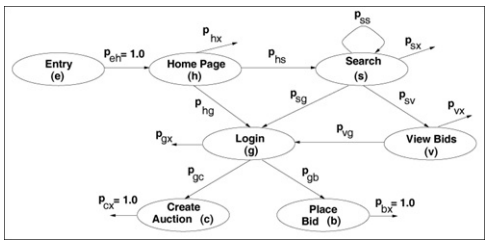
**Matriz de demandas D**: permite almacenar las demandas para cada componente del sistema por clase de operación.

**Vector N**: Especifica el número de instancias por componente

**Vector λ**: Almacena la tasa de llegada para cada clase de operación

**Parámetro γ:** carga total que llega al sistema.

**Grafo a usar para la práctica:**



Breve descripción:

-Este es un grafo que representa cualquier tipo de aplicación Web de manera general.

Nodo (e) -> representa al cliente entrando al sitio.

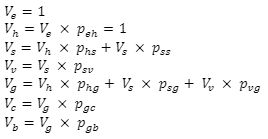
Arcos -> representan la probabilidad de transición directa de nodo a nodo.

Los nodos de salida (x) -> es el estado al que el cliente puede acceder desde cualquier estado i con probabilidad Pix.

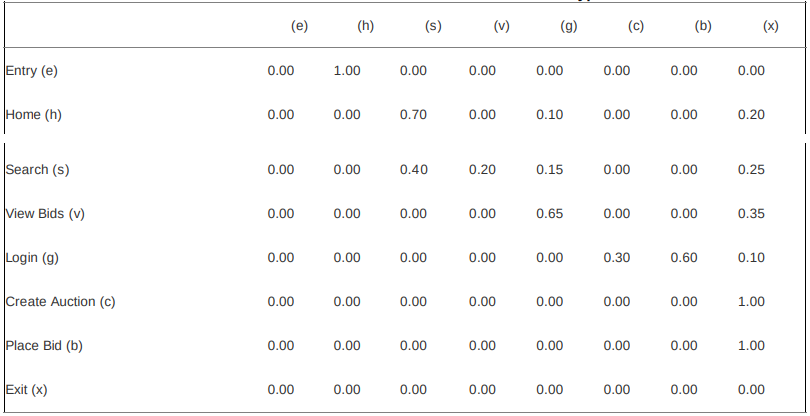
**Procedimiento:**

1 - Se obtienen las cargas de trabajo para cada tipo de clase de operación. Este proceso comienza obteniendo las visitas a cada estado realizadas por un cliente en una sesión. Se usan las tablas 1 y tabla 2 para obtener las probabilidades de transición de los estados.

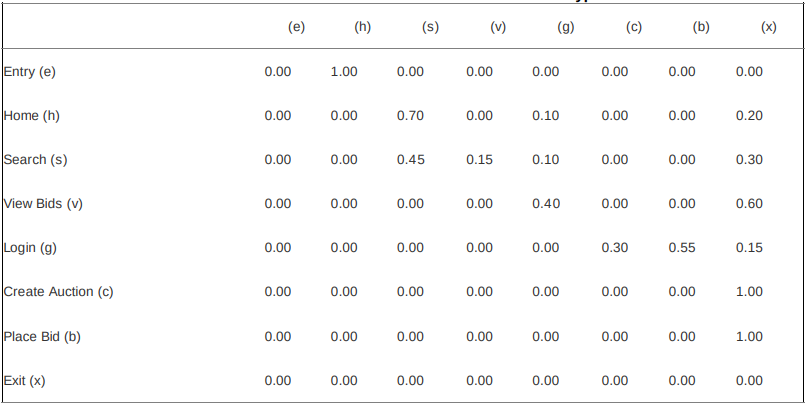
Formulas:



\*Formulas correspondientes al grafo anterior



**Tabla1**



**Tabla #2**

2 – Una vez resuelto el sistema lineal anterior y obtenidos los resultados de las visitas, la definición de **λ** como la carga total que entra al sistema y teniendo a fA y fB

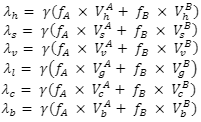
**λ** = 1

fA = 0.25

fB = 0.75

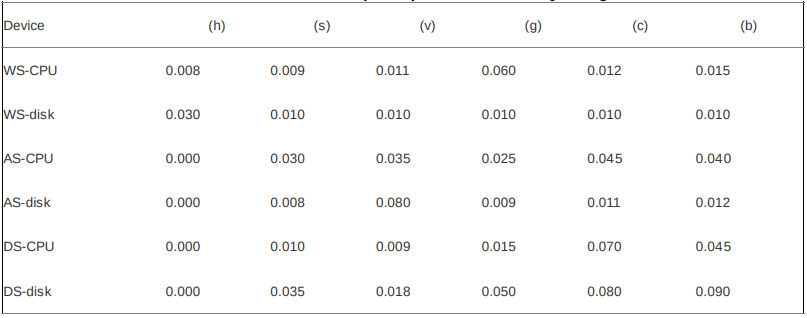
Se procede a calcular la carga para cada una de las anteriores clases.

**Formulas**:



3 – Se procede a calcular la utilidad para definir cuan útil es el modelo.

Se hace uso de la matriz de demandas dada en la Tabla#3



**Tabla #3**

**Formula**:



**\*para calcular las utilidades**



**\*para calcular los tiempos de respuesta**

Nota: hay que revisar todos los métodos de manera general, el método para calcular los tiempos de respuesta no creo que está muy correcto, aun cuando la fórmula es simple, no entiendo el cómo operan en el ejemplo, además va faltando poner el ejemplo concreto o los diferentes ejemplos con las tablas e ir analizando la variación de los resultados.