**Métodos cuantitativos**

**Dr. Raúl V. Ramírez Velarde**

**Proyecto de Simulación de Transmisión de Video en la Nube**

**Proyecto Final**

**1. Introducción**

Los servicios de video bajo demanda (VOD) son servicios en los que el usuario selecciona cuando va acceder al contenido, ya sea audio o video. El contenido es distribuido a los usuarios desde el servidor utilizando una red de datos por medio de la tecnología streaming. En Internet, como en Youtube y Netflix, la transmisión de video se realiza sobre HTTP, aunque hay otras opciones para v video distribuido en medios de transporte y en proyectos residenciales u hoteles. Los videos típicamente se almacenan en formato MPEG-4 o H.263, aunque hay una gran variedad de formatos. Los videos a distribuir se pueden almacenar en un sistema de almacenamiento totalmente operado por el proveedor de servicio, en un entorno de hosting distribuido a través de todas las áreas de servicio o en un servicio en la nube, como Netflix (Utiliza Amazon EC2 desde 2010). Se calcula que en 2016 el 51% de las ganancias en publicidad de medios provendrán del VOD basado en Internet o 34 mil millones de euros. El número de usuarios de VOD en el mundo llegará a miles de millones pronto. Sólo Netflix tiene 74 millones de subscriptores. En este proyecto se desea encontrar los parámetros operativos de un sistema de distribución de video que funciona en la **nube**.

**2. Descripción**

En este proyecto, obtendrás los datos de los videos a utilizar del sitio de Internet “MPEG-4 and H.263 Video Traces for Network Performance Evaluation” de la Universidad Técnica de Berlín, y se te proveerá de un archivo Excel que contiene una muestra de los datos de retraso de acceso a la nube Amazon EC2. Con esta información, determinarás la calidad de servicio que puedes ofrecer de acuerdo al número de usuarios simultáneos y los parámetros de operación, como tasa de servicio para cada cliente. El proyecto estará dividido en 4 etapas.

1. Análisis estadístico
2. Simulación
3. Modelación
4. Conclusiones

**2.1 Etapa 1: Análisis Estadístico**

1. **Análisis de Componentes Principales**

En este paso determinarás el comportamiento característico. La base de datos de trazas de video de la Universidad Tecnológica de Berlín contiene muchos videos y no hay forma de saber cuáles de ellos serían más populares en un cierto momento. Por eso, se desea crear una traza sintética que captures la mayor cantidad de la variabilidad de todos los videos. Es decir, un video sintético que capture el comportamiento de casi todos los demás video. Realiza un análisis de componentes principales. Esto es básicamente determinar los valores y los vectores característicos de la matriz de correlaciones. Para este análisis, las variables serán cada video, y las muestras cada uno de los marcos comprimidos. Es decir, crearás una matriz de la siguiente manera:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tamaño de marco en bytes | Jurassic Park | Silence Of The Lambs | Star Wars 4 | Mr. Bean |
| Marco 1 | 534 | 18059 | 919 | 600 |
| Marco 2 | 134 | 7274 | 302 | 144 |
| Marco 3 | 390 | 8825 | 432 | 1320 |
| Marco 4 | 1542 | 14940 | 1161 | 3888 |
| Marco 5 | 407 | 6809 | 284 | 885 |
| … | … | … | … | … |
| Marco 64,000 | 679 | 6325 | 2515 | 2619 |

Realiza el análisis de componentes principales (PCA). Puedes usar Scilab, R, Matlab o XLSTAT (y muchos otros más), aunque se te aconseja usar XLSTAT. Consulta en Youtube cómo realizar estos análisis en la herramienta que selecciones. Para XLSTAT, puedes usar las computadoras que están en 3-401, las 20 computadoras centrales tiene licencias de XLSTAT, o descarga XLSTAT con licencia de un mes. Una vez que realices el PCA, las coordenadas de las observaciones del primer componente principal (F1 Scores) será la base para el video característico (CVT).

Para recrear CVT, utiliza la siguiente fórmula:

Donde X el F1, o el primer componente principal, es la desviación estándar que se desea (puedes obtener la desviación estándar de todos), es la media que se desea y es el valor característico más grande de la matriz de correlaciones. Si desea agregar más componentes, entonces:

Ahora debes proceder al análisis estadístico.

1. **Determinación de la normalidad de los datos**

Usa el documento “Probability Plots to Compare Data with Distributions (Probability Plots)” para realizer un diagram de probabilidad y determiner (ver el final del documento) si el CVT es de cola pesada o cola ligera.

1. **Análisis estadístico descriptivo**

Usando el paquete de análisis estadístico obtén la media, la varianza, la desviación estándar, la simetría y la curtosis del CVT. Usando la curtosis, determina si el CVT tiene cola pesada. Realiza también un histograma (con líneas, no barras) del CVT. ¿A qué distribución de probabilidad se parece? ¿Se ve picuda con cola pesada o se ve más bien suave y bajita?

1. **Ajuste de distribuciones de probabilidad**

Ahora realiza un ajuste de distribución de probabilidad. Hay 3 tutoriales que puedes consultar. Estos son: “StatToolbox Tutorial”, “DifTool” y “StatToolbox-Allfit”. Identifica qué distribución de probabilidad describe mejor los datos. Prueba las siguientes distribuciones de probabilidad al menos:

1. Exponencial
2. Gamma
3. Weibull
4. Lognormal
5. Pareto

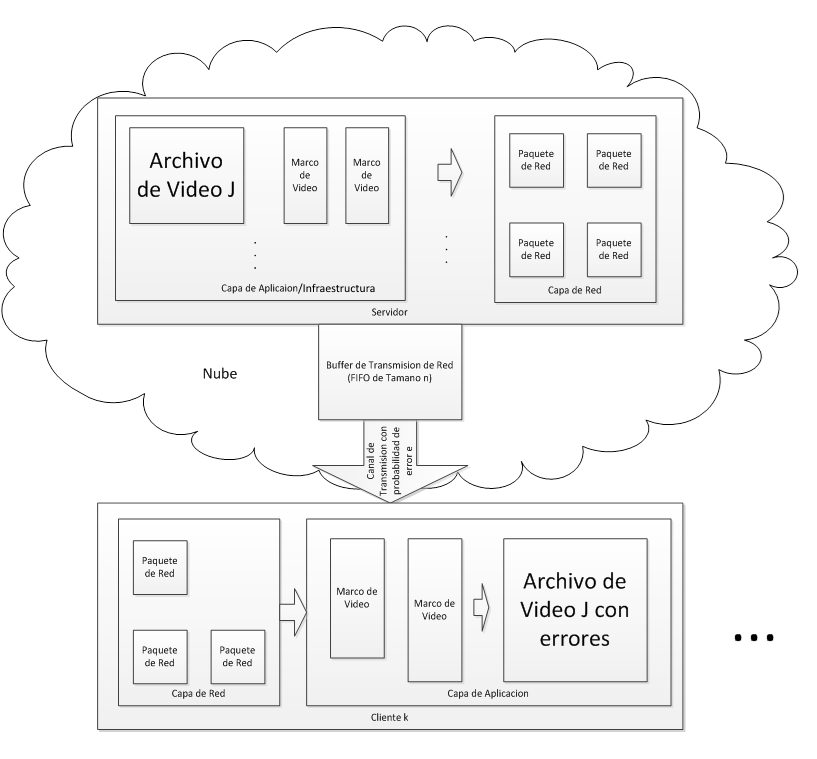
Realiza gráficas cuantil-cuantil (qqplot) con cada distribución de probabilidad y enlista los parámetros encontrados.

**2.2 Etapa 2: Simulación**

El objetivo principal de la simulación en encontrar **NURX** (Número de Usuarios, Utilización, Retraso y Rendimiento).

En este sistema, un usuario entra en el sistema de la nube y solicita un video particular. El sistema lee los segmentos que conforman el video uno por uno, de forma secuencial en la capa de aplicación la cual llamaremos de infraestructura. De ahí, utilizando una disciplina de primeras entrada primeras salidas, el sistema envía un cuadro completo de video hasta que llega a la capa de red en donde se dividen los marcos de video en paquetes de un determinado tamaño máximo (**MTU**). Esto implica que marcos de video de diferentes tamaños se subdividen en paquetes más pequeños para ser cargados en un buffer de salida de tamaño ***n*** y después viajar por la red interna de la nube y el internet hasta el usuario final.

El siguiente diagrama ilustra la situación. Hay un servidor en la nube que atiende a ***K*** clientes con archivos de video que pueden o no ser los mismos, pero que son tratados como distintos por el sistema (pues cada uno se sirve a diferente tiempo). Para que los paquetes de red viajen, antes pasan por un buffer el cual tiene un tamaño limitado **n** (medido en número de paquetes), el cual se comporta como una cola FIFO. Este además rechaza paquetes entrantes cuando ésta esté llena. Los paquetes viajan por un canal de red interna y externa a la nube que tiene error y pérdidas, por colisiones, interferencia, etc. La probabilidad de que un paquete que viaje por ese canal no llegue a su destino es ***e***, el cual es variable y dependiente de la cantidad de clientes k.



Los clientes nuevos llegan a razón de 1/***tp***. Cada video se reproduce a una tasa de **24 marcos** de video por segundo. Cuando se comienzan a enviar marcos de video a cada cliente, en la capa de infraestructura se dividen los marcos de video y se mandan a la capa de red forma round robin sin prioridad a razón máxima **r** (**1,000** marcos/seg). Así mismo salen de la capa de Red en forma de paquetes con disciplina FIFO hacia el buffer de transmisión. La comunicación entre cliente y la nube se considera que es mediante protocolo UDP, por lo que si un paquete no llega al cliente este se considera como perdido, no hay ACK. El ancho de banda del medio es de **4,500** paquetes de red por segundo (es decir 54 Mbps, Wifi). Esto quiere decir que se elimina un paquete de alguno de los buffer cada **222** microsegundos. Adicionalmente se agrega un retraso ***d*** que tiene la nube al buscar los marcos de video y el tiempo de red interna de la nube y del internet hasta el usuario final. Este retraso ***d*** es variable y depende de la carga de la nube.

Puede darse que exista saturación en el servidor en la nube, donde se manden paquetes desde el servidor de la nube a la capa de red y que estos no puedan ser enviados porque la cola este llena y se pierdan segmentos de marcos de video. La probabilidad de que esto ocurra es ***G***.

Asimismo, queremos que el retraso total máximo (***d*** + tiempo en cola) sea menor a un tiempo ***U.*** Puede ser que la nube o la cola de la red tengan tiempos de retraso muy altos en ese momento y que se exceda el tiempo máximo ***U.***

Para cada una de las siguientes situaciones determina NURX, esto es:

1. Número medio de tareas en el búfer
2. Utilización de la red Wifi
3. Tiempo promedio de residencia en el búfer
4. Rendimiento promedio de la red Wifi
5. Probabilidad de saturación del búfer

**Lo que se desea saber es lo siguiente:**

1. Dada un cierto tamaño de buffer **n**=50, 100, 250, 500 (tamaño máximo de cola en paquetes), una probabilidad de error **e**=0.001\***k**, suponiendo la cantidad de usuarios **k** = 5, 10, 15, 20 y si queremos que la probabilidad de saturación **G** sea menor o igual a 0.05, 0.01 o 0.001,¿cuál es el tamaño **n** (en paquetes) que debe tener el buffer de transmisión?
2. ¿Existe una cantidad máxima de usuarios que se puede atender incluso si no hay errores en la red?

Algunos parámetros operativos a considerar son los siguientes:

Los paquetes en el buffer de tamano **n** pueden ser desde 20 hasta **MTU** bytes

**MTU** = 1500 bytes (tamaño máximo de paquete)

**t**=222 microsegundos (tiempo entre salidas de paquete de búfer)

**r** = 1,000 marcos/segundo (tasa de transferencia de la nube a la red)

**tp**= 2 segundos

=Tasa de entrada de paquetes= (tasa variables)

=Tasa de salida de la cola de red (5,400 paq/seg)

Los cuadros de video que se pueden utilizar son los que se encuentran en

<http://www-tkn.ee.tu-berlin.de/research/trace/ltvt.html>

Los archivos .dat contienen la información del tamaño del marco de video de toda la secuencia en bytes. Para obtener los tamaños de marco que se envían, utilice la versión terse de los frame traces disponibles. Por ejemplo, el primero de los enlaces de esta película y calidad escogidos ([Terse\_Jurassic.dat](http://www-tkn.ee.tu-berlin.de/research/trace/pics/FrameTrace/mp4/Terse_Jurassic.dat)) encontrado en <http://www-tkn.ee.tu-berlin.de/research/trace/pics/FrameTrace/mp4/indexaa60.html>

Use los archivos .dat de la película que quiera en la calidad que desee. En este caso use el mismo archivo de marcos para cada cliente, aunque cada uno empezara a servir a distinto tiempo. Use 2000 marcos continuos seleccionados de una sección que no sea el inicio de los .dat

Para los tiempos de retraso **d** de la nube use los datos de retraso ejemplo (de Amazon) proporcionados por el Profesor. Considere los renglones como el tiempo de retraso de la nube con una resolución en microsegundos. Es decir, renglón 1 es tiempo 0, renglón 2 es el retraso de la nube en el siguiente microsegundo, etc. Las unidades de los retrasos en la hoja de Excel (AmazonS3\_delays) están en microsegundos.

Muestre tablas y graficas de resumen con sus resultados intermedios y responda las preguntas que se plantearon en un reporte.

Entregue el reporte, código fuente y archivos necesarios en un .ZIP

**Notas aclaratorias:**

Cada marco de video genera varios paquetes de red. Por ejemplo, si el marco de video es de 3,831 bytes, este generará 3 paquetes de red; dos de tamaño 1,500 bytes (MTU) y uno de 831 bytes (3,831=1,500x2+831)

En la capa de infraestructura en la nube, los marcos de video son divididos en paquetes de red y enviados a la capa de red a una tasa de 1,000 clientes/seg. Suponiendo 3 paquetes de red por marco de video, eso da una tasa máxima de 1,000x3=3,000 marcos de video por segundo. Sin embargo, si temporalmente se tiene marcos de video grande que generen 5 paquetes de red, la tasa de la capa de infraestructura es de 1,000x5=5,000, superior a la tasa de salida que provee la red Wifi. Cada 0.001 segundos se toma un marco de un cliente y se segmenta en 1, 2, 3 … m, paquetes de red y se meten en la fila. La capacidad de la red Wifi, o sea de la tasa de salida de la fila, es de =4,500 paquetes por seg para TODOS los clientes. Es una tasa fija.

Vamos a suponer que hay 3 clientes

El siguiente marco del cliente uno es de 3,831 bytes, o sea 3 paquetes

El siguiente marco del segundo cliente es de 4,615 bytes, o sea 4 paquetes

Y para el tercero, un marco de 850 bytes o sea un paquete.

Usando round robin:

* En el tiempo 0.001 se meten al buffer los tres paquetes del usuario 1.
* En el tiempo 0.002 se meten los 4 paquetes del usuario 2.
* Y en el tiempo 0.003 el paquete del usuario 3.

¿Qué pasa después de esto? La capa de infraestructura reinicia con el cliente 1. La tasa de la capa de infraestructura en clientes por segundo es constante y en paquetes por segundo es constante. Esta tasa no se ve afectada por el número de clientes. Si sólo hay un cliente, a este se le atiende 1,000 veces en un segundo.

Mientras tanto, cada 0.000222 segundos sale un paquete del buffer. El retraso **d** (por ejemplo del excel de Amazon) se suma después de que los paquetes salen de la cola. O sea el tiempo de residencia en la cola + d es el retraso total. Ese tiempo está en microsegundos.

Por ejemplo, el primer tiempo de retraso de Amazon es de 248 microsegundos. Vamos a suponer que hay 60 clientes. Esto solo afecta en el tamaño promedio de los paquetes. Supongamos que en este instante, todos los clientes generan 4 paquetes por marco de video. La nube debe de mandar aproximadamente 4,000 paquetes por segundo a la capa de red (1,000 clientes/seg x 4 paq/cliente). Si la capa de red tiene una capacidad de 4,500, el tiempo promedio en cola sería segundos, o sea 2 milisegundos. El tiempo total de retraso por paquete de red sería 0.002248 segundos, o 2.248 milisegundos o 2,248 microsegundos.

Para simplificar la simulación, el error y el retraso se consideran a nivel de paquete de red. Por cada paquete que sale de la cola, se debe tirar un número aleatorio. Si este número es menor que la probabilidad **e** de error en paquete, el paquete se pierde y el marco de video no se completa. Si la cola está llena, también se pierde el paquete.

**2.3 Etapa 3: Modelación**

En esta fase del proyecto deberás determinar qué modelo de los visto en clase o presentados por tus compañeros se ajusta más a los resultados obtenidos. Se te pido que hagas el análisis en dos pasos.

1. Leyes Operativas
2. Modelación por Teoría de Colas
3. **Leyes Operativas**

Haz un análisis de valor medio utilizando las fórmulas de leyes operativas. Determina NURX, esto es:

1. Número medio de tareas en el búfer
2. Utilización de la red Wifi
3. Tiempo promedio de residencia en el búfer
4. Rendimiento promedio de la red Wifi
5. ¿Son congruentes estos resultados con los obtenidos en la simulación?
6. **Modelación por Teoría de Colas**

Utiliza las fórmulas de los diversos modelos y establece cuál es el que mejor se ajusta al sistema que has modelado. Los modelos conocidos son:

1. Colas Markovianas
   * M/M/1
   * M/M/k
2. Colas Generales
   * M/G/1
   * M/G/k
   * M/G/∞
   * M/G/k/k
   * GI/G/k
3. Mas Modelos
   * Cloud Perfectamente Elástica
   * Llegadas con Raje

Selecciona al menos 4 modelos. Para cada uno de estos modelos determina NURX, esto es:

1. Número medio de tareas en el búfer
2. Utilización de la red Wifi
3. Tiempo promedio de residencia en el búfer
4. Rendimiento promedio de la red Wifi
5. Probabilidad de saturación del búfer

***Si desarrollas tu propio modelo utilizando procesos markovianos, aunque no obtenga mejores resultados que los modelos ya establecidos tendrás puntos extras.***

**2.4 Etapa 4: Conclusiones**

Primero, enlista todo lo que te llamó la atención en la realización de este proyecto y reflexiona acerca de esos aspectos. Incluye hechos, acciones y pensamientos. Ahora piensa acerca de tus responsabilidades como ingeniero de cómputo, es decir, tus responsabilidades hacia quienes financian el proyecto poniendo en riesgo su capital y acerca de los usuarios que también pondrán capital al pagar los servicios y que esperan recibir cierta calidad. Los usuarios quieren una inversión mayor para tener menores tiempos de espera y mejor rendimiento, mientras que el inversionista prefiere una inversión menor para mejorar el retorno de capital. ¿Qué actitud debes de tomar tú? ¿Debes permitir el sobre diseño del sistema para favorecer a los clientes? Debes permitir el subdiseño del sistema para favorecer a los inversionistas? ¿Crees que existen alternativas al diseño que se te proporcionó que sean más eficientes?

Desde el punto de vista de ciudadanía, ¿cómo crees que una correcta modelización del proyecto ayude a tener una mejor comunidad, una mejor ciudad, un mejor país?

Y desde el punto de vista de desarrollo sustentable, ¿cómo afecta al medio ambiente una correcta modelización del sistema propuesto?

**3. Rubrica**

|  |  |  |  |
| --- | --- | --- | --- |
| Item | Beginer (50%) | Competent (80%) | Proficient (100%) |
| Statistical Analysis  (30%) | Student is able to create a PCA with XLSTAT or similar. Selection of data from original source is completely justified | Characteristic video trace is correctly constructed. Student is able to carry out descriptive statistic analysis of the CVT, also Normality plot and to fit different probability distributions. **ALL PLOTS HAVE TITLE, AXIS VARIBLES IDENTIFIED, AND THE UNITS OF EACH AXIS.** | Several PCAs are carried out to determine the correct collection of videos that achieves maxium variability on the first 3 PCs. Statistical analysis is suported by plots. The best pdf is selected. **Student is able support conclusions with plots and facts clearly described in her report.** |
| Simulation  (40%) | Is able to substitute exponential random variables for variables read from a file in Excel or programming language. Student is able to assign a video to a simulation client. Student is able to include Cloud delays and transmission errors in simulation | The simulation computes statistics for usage, throughput, tasks on queue and system wait time. Student is able to compare simulation results. Student is able to determine which operational configuration is best considering delay and saturation probability limits.. She can conclude how many users can be serviced simultaneously. **ALL PLOTS HAVE TITLE, AXIS VARIBLES IDENTIFIED, AND THE UNITS OF EACH AXIS.** | The simulation is able track more events than just time of departure and time of arrival (por example syncronic statisitics, rejection, etc.)**.** Student justifies the comparison using graphs, plots, excel tables and running different scenarios for different values. **Student is able support conclusions with plots and facts clearly described in her report.** |
| Modelling  (20%) | Mean value analysis is effectively carried out | Several queuing models are tried and results presented. **ALL PLOTS HAVE TITLE, AXIS VARIBLES IDENTIFIED, AND THE UNITS OF EACH AXIS.** | Student determines a selection criteria to choose a model. The best model is selected. If no model can describe the simulation an explanation is offered. **Student is able support conclusions with plots and facts clearly described in her report.** |
| Conclusions (10%) | Student is able to draw technical conclusions by using facts, actions and observations relating the three previous phases of the project | Student is able to draw ethical conclusions from her work. Student is able to draw citizenship and environmental conclusions from her work | Conclusions are enhaced with researh on ethical, citizenship and environmental aspects of the project |