Utilidad de los flujos NetFlow de RedIRIS para análisis de una red académica

Utility flows NetFlow RedIRIS for analysis of an academic network

J. L. García-Dorado, J. E. López, J. Aracil, V. López, J. A. Hernández, S. López-Buedo y L. de Pedro

Resumen

La tecnología Netflow se encuentra actualmente desplegada en la mayoría de los routers de las redes comerciales. Dicha tecnología permite registrar los flujos que atraviesan la red, contando el número de bytes y paquetes que se transmiten entre dos equipos, lo cual puede ser útil para diversas aplicaciones.

Este trabajo muestra la utilidad y aplicabilidad a la gestión de los flujos de red. En concreto, se presenta cómo los registros de flujos de red que captura RedIRIS en los nodos autonómicos pueden resultar de interés para monitorizar redes y como herramienta de obtención de medidas de la red para su posterior análisis con múltiples fines. Para ello, se ha implementado una aplicación que accede a estos registros, previamente procesados, y muestra de forma sencilla un conjunto de medidas y estadísticas para la mayoría de las universidades conectadas a RedIRIS.

Palabras clave: monitorización de red, NetFlow, análisis de tráfico.

Summary

Netflow technology is currently deployed in most commercial network routers. This technology allows the record of flows traversing a network, counting the number of bytes and packets that are transmitted between two computers, which can be useful for diverse applications.

This work shows the utility and applicability of flow-based network management. Specifically, it is shown how the network flow records captured by RedIRIS at the autonomic nodes can be interesting to monitor networks and as a tool to obtain measurements of the network for a later analysis with multiple purposes. For this, an application has been implemented that accesses these records, previously processed, and shows in an easy way a set of measurements and statistics for most of the universities connected to RedIRIS.

Keywords: network monitoring, NetFlow, traffic Analysis.

La tecnología Netflow se encuentra actualmente desplegada en la mayoría de los routers de las redes comerciales

1. Introducción

Las posibilidades que permite la tecnología NetFlow[1], definida inicialmente por Cisco[2], son muy importantes en áreas muy diversas que incluyen la monitorización, la predicción de ataques o la detección de intrusos[3]. Además, también permiten determinar las razones por las que se genera un cuello de botella en una red, el lugar donde este se produce y, en general, cualquier tarea de análisis de la red. Son, por tanto, una herramienta muy potente que puede resultar de gran utilidad en la gestión de redes.

Las posibilidades que permite la tecnología NetFlow son muy importantes en áreas muy diversas que incluyen la monitorización, la predicción de ataques o la detección de intrusos

Este trabajo pretende mostrar cómo se pueden capturar los flujos de red que generan los routers autonómicos de los que dispone RedIRIS[4] y cómo se puede extraer de ellos medidas y estadísticas de la red. En primer lugar, se presentará una pequeña descripción de los flujos de red y se comentará el modo en que los routers crean un registro por cada flujo que encaminan. A continuación se mostrará la infraestructura necesaria para capturar y almacenar estos registros. Igualmente, se comentarán las dificultadas encontradas a la hora de trabajar con ellos. Estas tienen que ver, principalmente, con las limitaciones hardware de los routers, pero también con el proceso de captura, ya que éste almacena los datos como una sola entidad sin diferenciar su procedencia. Finalmente se presentará una aplicación web, implementada para facilitar el acceso a los datos y su visualización. En concreto, esta aplicación mostrará alguna de las métricas de red que se pueden obtener utilizando los flujos de red. Esta aplicación nos permitirá, también, validar la corrección de todo el proceso. Finalmente se comentarán las conclusiones y se planteará el trabajo futuro.

2. Flujos de red (NetFlow)

Entendemos por flujos de red, más conocidos como NetFlows (denominación de Cisco), como el conjunto de paquetes sucesivos que comparten protocolo, direcciones IP y puertos (origen y destino). Respecto a esta definición genérica, se pueden añadir más campos que determinen la unicidad de un flujo. En concreto, Cisco añadió otra serie de identificadores como son el tipo de servicio (ToS) o las interfaces de red. Este conjunto de identificadores permiten identificar de forma única a un flujo.
En un primer momento, Cisco pensó que mantener una serie de registros con las estadísticas de cada flujo podría ser de gran interés para facilitar el proceso de encaminamiento y para ello implementó la tecnología NetFlow. Rápidamente se le han encontrado otras utilidades como ya hemos comentado en la introducción.
Estas estadísticas consisten, básicamente, en los tiempos de inicio y fin del flujo, número de paquetes y bytes, interfaces de salida y entrada, direcciones IP y puertos (origen y destino), etc.
De esta forma, cuando un router consideraba que un flujo había finalizado, lo eliminaba de la memoria interna y permitía exportar su información. Esta información es de gran interés para el análisis del tráfico de cualquier red ya sea en el mismo momento de exportarla o para analizarla con posterioridad tras su almacenamiento. Obviamente, debe definirse qué se entiende por fin de un flujo. En general se considera que un flujo ha finalizado cuando no se observa tráfico durante 15 segundos, cuando está activo durante más de 30 minutos, cuando se encuentra una bandera de fin de conexión o, simplemente, cuando el router se queda sin recursos. De hecho, la carga que exige actualizar cada entrada y celda de la tabla de flujos por cada paquete que llega a un router, es decir, el análisis de las cabeceras de todos los paquetes entrantes, es, frecuentemente, excesivo para las capacidades actuales de los routers en cuanto a acceso a memoria y coste computacional.
En este sentido, la solución implementada para evadir estas limitaciones consiste en analizar tan sólo un porcentaje de los paquetes que le llegan a los routers, de modo que se reduzca el trabajo a realizar[5]. Esto se traduce en que, de forma aleatoria y según una tasa de muestreo, algunos paquetes no son analizados y su información no queda reflejada en las estadísticas de los registros NetFlow. Esto ocasiona una imprecisión en los registros de los flujos. Métricas como el tiempo de inicio y fin pueden verse alteradas al no ser tenido en consideración el primer o el último paquete. En cuanto a estadísticas como el número de paquetes y bytes, se verán reducidos de forma inversamente proporcional a la tasa de muestreo. Incluso, puede darse el caso que un router no llegue a detectar un flujo si éste es lo suficientemente pequeño. Es decir, tras el submuestreo los datos no podrán ser recuperados de forma exacta pero, como veremos en este trabajo, la precisión perdida no es de gran relevancia en muchas métricas.

Cisco pensó que mantener una serie de registros con las estadísticas de cada flujo podría ser de gran interés para facilitar el proceso de encaminamiento y para ello implementó la tecnología NetFlow

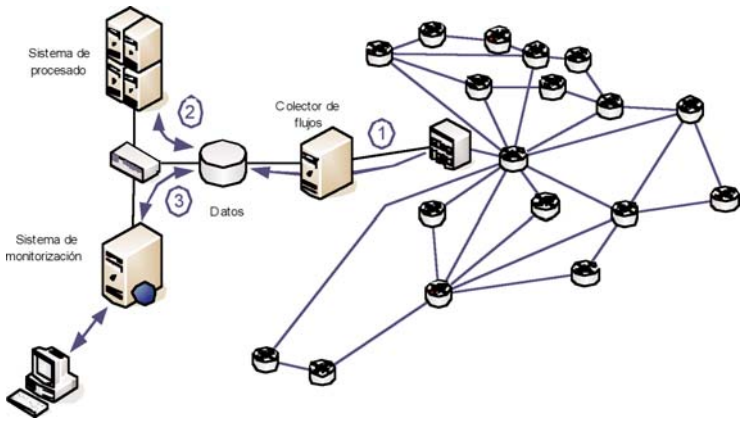
Tras el submuestreo los datos no podrán ser recuperados de forma exacta pero, como veremos en este trabajo, la precisión perdida no es de gran relevancia en muchas métricas

3. Infraestructura de medida, almacenamiento y análisis

La figura 1 muestra, de forma simplificada, la topología de RedIRIS. En concreto, a la derecha, se muestran los principales nodos (routers) que la forman y sus respectivos enlaces. La parte izquierda de la figura muestra la infraestructura emplazada en la Universidad Autónoma de Madrid (UAM) para capturar, almacenar, analizar y mostrar estadísticas del tráfico generado por muchas de las universidad que están conectadas a RedIRIS. Más en concreto, cada uno de estos nodos ha estado enviando, desde abril de 2007, sus registros NetFlow al colector de flujos, fase 1 en la figura. Toda esta información ha sido almacena utilizando la herramienta Flow-tools[6]. El tamaño de la información almacenada, hasta octubre de 2007, ha sido de unos 2TB a una tasa media de llegada de 2Mbps.
A continuación, y una vez convenientemente almacenados los datos, se ha procedido al análisis, representado en la figura 1 como la fase 2. En particular se han calculado estadísticas sobre los puertos y direcciones IP más activas, el ancho de banda consumido y la hora más cargada.

Finalmente, se dejan los datos a disposición del sistema de monitorización que permite acceder a ellos mediante su descarga o visualización de forma cómoda mediante un navegador web (fase 3).

FIGURA 1. TOPOLOGÍA DE RedIRIS E INFRAESTRUCTURA DE MEDIDA, ALMACENAMIENTO Y ANÁLISIS



4. Problemática

Cuando se desea analizar los registros NetFlow surgen una serie de problemas que este trabajo pretende analizar

En la práctica, cuando se desea analizar los registros NetFlow surgen una serie de problemas que este trabajo pretende analizar. Son dos los principales problemas a subsanar: primero, los datos, como se comentó en la introducción, están muestreados. Y segundo, los datos que se reciben en el colector de flujos engloban los registros de todos los routers de RedIRIS. Es decir, los registros de todos los nodos mostrados en la figura 1 vienen mezclados. Además, en el mismo sentido, es bien sabido que RedIRIS da servicio no sólo a universidades (centros objetos de estudio en este trabajo) sino también a hospitales, institutos, observatorios, etc. En consecuencia, parte del total de registros recibidos hacen referencia a tráfico que no se desea analizar.

Para solucionar el primer problema, basta con multiplicar el número de paquetes que indica el registro del flujo muestreado por el inverso de la tasa de muestreo. Obviamente, la información no será igual de precisa, pero como veremos las diferencias serán de poca importancia en las métricas bajo análisis. Con respecto al segundo problema, se deben agrupar los flujos por router y por universidad. RedIRIS nos facilitó los rangos IP de la mayoría de las universidades españolas para realizar este proceso, así como las direcciones IP de los routers que nos han estado exportando los registros. Esta información es suficiente para filtrar, del total de la información recibida, los registros que hacen referencia al tráfico que generó cada universidad de forma independiente del resto de instituciones.

5. Herramienta

Se implementó una herramienta para acceder fácilmente a los datos previamente analizados y que además nos permitirá estimar la corrección de todo el proceso

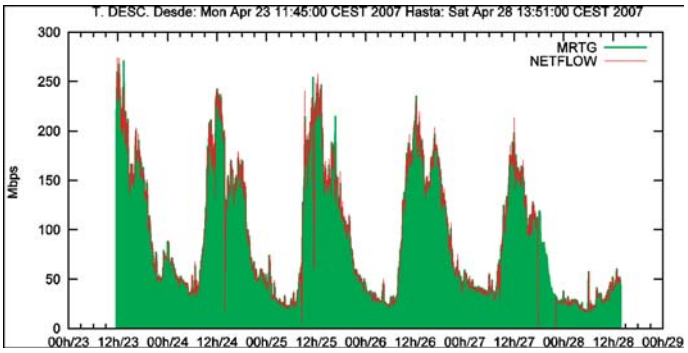
Finalmente, se implementó una herramienta para acceder fácilmente a los datos previamente analizados (fase 3 de la figura 1) y que además nos permitirá estimar la corrección de todo el proceso. Esta aplicación, accesible mediante una interfaz web, permite visualizar el ancho de banda consumido por muchas de las universidades que forman redIRIS en intervalos de 5 minutos. También permite visualizar cuándo se produjo la hora más cargada y el tráfico medio en ésta, así como aquellos puertos y direcciones IP (convenientemente anonimizados) que han producido la mayor parte del

tráfico. Los puertos, por su parte, se pueden ver desglosados según sean bien conocidos o no, o si son alguno de los puertos más típicos (web, correo, DNS, etc.) Del mismo modo, la aplicación, permite descargarse estos datos en un fichero plano para su posterior análisis.

Una forma fácil de comprobar la corrección de los resultados de esta nueva aplicación es compararlos con el ancho de banda consumido estimado mediante la herramienta MRTG[7]. Sin embargo esta comparación sólo es posible en aquellas universidades donde, por estar conectadas a RedIRIS directamente, los valores MRTG reflejen el tráfico de la universidad sin tráfico cruzado. Es decir sin que el router encamine ningún tráfico de otros centros salvo el propio de la universidad. Los resultados fueron completamente satisfactorios. A modo de ejemplo, la figura 2 muestra el ancho de banda durante seis días aproximadamente en una universidad junto al tráfico estimado mediante MRTG. Como puede comprobarse las dos gráficas están superpuestas (línea roja para los flujos, fondo verde para MRTG), lo que indica que todo el proceso seguido fue correcto[8]. Por otro lado puede comprobarse cómo en algún momento de los seis días que representa esta última figura, se dejaron de recibir datos a causa de algún problema en los routers o en la infraestructura de captura.

Una forma fácil de comprobar la corrección de los resultados de esta nueva aplicación es compararlos con el ancho de banda consumido estimado mediante la herramienta MRTG

FIGURA 2. ANCHO DE BANDA CONSUMIDO POR UNA UNIVERSIDAD DURANTE SEIS DIAS TAL COMO LO MUESTRA LA APLICACIÓN WEB. SE MUESTRA TANTO EL TRÁFICO CALCULADO A PARTIR DE NetFlow COMO MEDIANTE MRTG.



6. Conclusiones y trabajo futuro

Las medias obtenidas mediante los flujos de red están siendo ya utilizadas en el proyecto DIOR

Este trabajo ha mostrado que las utilidades de los flujos de red son importantes y aplicables a muchos campos de la gestión de red. En concreto, se ha comprobado cómo los registros de flujos de red que captura RedIRIS en los nodos autonómicos pueden resultar de interés para monitorizar redes y como herramienta de obtención de medidas de la red para su posterior análisis.

Igualmente, se ha implementado una aplicación que accede a estos registros, previamente procesados, y muestra de forma sencilla un conjunto de medidas y estadísticas para la mayoría de las universidades conectadas a RedIRIS. Por otro lado, con intención de evaluar la corrección de la aplicación, se han comparado estas medidas con otros obtenidos de forma diferente, obteniendo los mismos resultados. De este modo se comprueba que los problemas, presentados a lo largo de este trabajo, pueden ser efectivamente superados.

En la práctica, las medias obtenidas mediante los flujos de red están siendo ya utilizadas en el proyecto DIOR (Dimensionado de redes IP y redes ópticas: aplicación a la red académica española RedIRIS)[9]. Este proyecto conjunto entre la Universidad Autónoma de Madrid y RedIRIS pretende relacionar las características que definen una universidad (esto es, su tamaño, las carreras disponibles, capacidad del acceso a RedIRIS, filtrado de aplicaciones o webs, etc.) y el tráfico que ésta genera. El objetivo final es ser capaces de encontrar reglas que permitan dimensionar redes de la forma más justa y objetiva posible.

Las mayores dificultades que se están encontrando a la hora de caracterizar la red de todas las universidades que forman RedIRIS son las relativas a obtener características relacionadas con la configuración de la red de cada universidad, pues no existe ninguna documentación centralizada. En general no resulta fácil determinar las políticas de filtrado que sigue cada universidad, la accesibilidad a Internet (esto es, filtrado de tráfico P2P y webs no permitidas) o la facilidades que tienen los alumnos para acceder a aulas con conexión a Internet.

Referencias

[1] *IP Flow Information Export. http://www.ietf.org/html.charters/ipfix-charter.html*
[2] *Introduction to Cisco IOS NetFlow. http://www.cisco.com/warp/public/732/Tech/NetFlow/.*
[3] *Francisco Monserrat, "Tecnologías de flujos en Red (NetFlow) para la detección de intrusiones y análisis forense", Jornadas Técnicas de RedIRIS. Nov. 2003.*
[4] *RedIRIS – Red española de I+D, http://www.rediris.es/*
[5] *C. Estan, G. Varghese, "New directions in traffic measurement and accounting: Focusing on the elephants, ignoring the mice", ACM Transactions on Computer Systems, Aug. 2003.*
[6] *flow-tools, http://www.splintered.net/sw/flow-tools/*
[7] *Tobi Oetiker's MRTG - The Multi Router Traffic Grapher, http://loss.oetiker.ch/mrtg/*
[8] *A. Feldmann, A.G. Greenberg, C. Lund, N. Reingold, J. Rexford, F. True, "Deriving traffic demands for operational IP networks: methodology and experience", IEEE/ACM Transactions on Networking, vol. 9, pp. 265-280, Jun. 2001.*
[9] *DIOR Project, http://www.ii.uam.es/~networking/projects/DIOR*

No resulta fácil determinar las políticas de filtrado que sigue cada universidad, la accesibilidad a Internet o las facilidades que tienen los alumnos para acceder a aulas con conexión a Internet

J. L. García-Dorado
J. E. López de Vergara
J. Aracil
V. López
J. A. Hernández,
S. López-Buedo
L. de Pedro
(proyecto.dior@uam.es)

Networking Research Group
Universidad Autónoma de Madrid