Good afternoon professors! It is great honor to show my final project for you. My name is Cheng Qian and my project title is User Behavior Analysis based on Deep Packet Inspection.

First of all, let me introduce the purpose and background of my project.

In this age filled with tons of information, the data became the most valuable source, especially those data specifically indicate every individual’s taste and habit. The Internet company (such as google, facebook, alibaba) pay much more attention to the subject of user behaviour analysis, because it can be applied in several aspects like accurate advertisement, push notifications, intelligent recommendation, and so on. In this way, those companies all increase the daily active user of their service. So we can see that the fact is all the user behavior data are going to the Internet service company.

This fact arouses me of my idea “what if we adopt it for network carrier (like AT&T, china mobile, china unicom)”. As the network carrier, they deliver the tube. Theoretically, none of the service companies could give more comprehensive data as the carrier do. Furthermore, under the rigorous control by the government, the database will become more secure and robust.

Then I turned to Deep Packet Inspection. It is a technique based on the Application layer. The characteristic of reading the load content inside of IP packets make DPI a possible way to obtain large amount of user behavior information. More importantly, it identifies protocol type through specific content feature, instead of only due to the port number, which make this technique much more reliable.

In the second part, I will show one of the two achievements: nDPI platform.

There are a lot of kinds of user behavior. My project mainly focused on those related to video websites. As you can see, users are using their devices to visit different video websites to watch the videos. The nDPI 1.6 platform was deployed at the switch of the lab network. So that it can inspect all the data flows through these links.

Then I will briefly describe the basic execution flow of the platform. Firstly, the platform will initialize the program. Then the system will assign a series of threads to circularly catch packets through the network card. After that, the platform will analyze the data packet from the lower layer to the upper layer, which concludes the Data link layer, Network layer, Transport layer and Application layer. For each layer, there are specific principles.

In Data link layer and Network layer, the system will first store the information of data link layer in “ndpi\_ethr” and set offset and data packet type. Then it will judge whether it is vlan\MPLS or PPPoE and store information of network layer in “iph”. Finally, the system will check if there exist GTP tunnel protocol. After all these process, the system will judge whether it is based on IP or other protocols, and obtain source/destination IP address, and protocol type.

In transport layer, first it will unpack the packets and obtain lower IP upper IP and protocol type, then assign tcp or udp according to different protocol number and obtain the lower port/upper port. After that the system calculates “idx” based on the five-elements set to identify each data flow. And based on whether there have been existed a binary linked list at this position, the system will decide whether to insert a new session into linked list. Finally, return the information of transport layer.

Finally the system will analyze protocol in the Application layer, which is the most important function in protocol analysis. It will firstly initialize the flow->packet constructor. Then the system will judge the position of this packet, for example, the three handshaking state of TCP connection establishment. After that, the constructor “ndpi\_selection\_packet” will be set, which is used to store the information of lower four layers. Finally, the system will call function “guessed\_protocol\_id” and “check \_ndpi\_flow\_func()” to inspect the protocol type in the application layer. The packets will be distributed to different interfaces according to its protocol in the transport layer (TCP\UDP\neither). The system will judge the protocol type by circularly call the discrimination function until inspection successfully, or set to UNKNOWN. The whole execution flow of the inspection platform is finished.

You can real-time grasp the Internet packets or give a pcap file to the system, the system will finally return statistical result of the packets in this pcap file.

In the third part, here come the other main achievement: front-end system.

I use regular expression to accurately identify and extract user behavior information keyword. Regular expression is a series of characters that define a search sequence pattern. It is mainly used for pattern matching string. For example, this figure shows the HTML code of Chinese biggest video website, Youku. In this HTML page, the main feature mainly concentrated in head part, the format of the header was summarized as this figure. Then the regular expression that used to extract the user key word and title can be showed as the code below.

I made a survey on six dominating video website’s HTML code and summarized their format into Regular Expression. Then I filtered and stored the user behaviour related factors (source IP address, destination IP address, video title and timestamp) into a MySQL database. (Because the object of my research is Chinese people, most of the video title is in Chinese). The database was maintained at localhost and can be connected to the local server.

Finally, a primary analysis front-end platform was programmed for querying every user's access behaviour details and regional integrated situation. Webpage was selected as a tool of my front-end query system because of its dynamism and transparency. The front-end query system was deployed on Apache Tomcat8.0. The main page was divided into two parts. The upper part is used to let the user set any conditions while querying, which include the query method, log ID, source IP address, key word or phrase of video and time range. The two buttons under the conditions are used to submit the form and reset the form. As soon as the submit button was pressed, the system will make a query to the data base and return the result. The lower part is used to display the query result and the statistical number of eligible logs. There are three query method : “Default”, “Find user” and “Find video”. I will show the details in the next demonstration part.

This is the whole schedule of my project and I have achieved each milestones from deployment, analysis, storage to front-end query system.

Now I will begin the demo part!