Balanced Scheduling Method for Big Data of Network Traffic Based on Set-pair Analysis Strategy

Yijie Yang,
Statistics, College of Science, NC State University, Raleigh, 276095,
yyang54@ncsu.edu

Abstract: The continuous surge in the total amount of big data of network traffic leads to excessive data transmission load on the network channel. To effectively alleviate this problem, a scheduling method for regulating big data load is designed and proposed. The link discovery is applied to perform set-pair analysis of network topology awareness; the cloud computing algorithm is used to perform calculations for data transmission; based on the measurement results, traffic routing is controlled and measured to achieve link analysis and reorganization, as well as completing balanced scheduling of traffic big data, which proves that the method can effectively alleviate the big data scheduling load problem and has an important role for research in network optimization fields.

1. Introduction
In the context of the rapid development of modern Internet technology, the total amount of China's network optical fiber data is increasing, and the concept of big data is emerging rapidly as well along with the development of the Internet and data transmission. The massive data flow behind it has caused a series of resource congestion while bringing convenience to people in data applications. It is undeniable that the popularity of modern network technology and data terminals have increased the amount of network data by countless, and the problem of domestic network user traffic balance is also showing. Although domestic network technology enterprises have begun to actively expand network channels and relevant data transmission services, from the macro perspective of network development, the load of China's network traffic control is still increasing, and the balance of data scheduling still faces certain problems. Balanced scheduling technology of big data traffic load can be seen as a series of measures taken to avoid network congestion caused by excessive network data load. The current balanced scheduling method of domestic network traffic big data mainly takes the reorganization of the channel as the core, with network traffic control link optimization strategy, or directly adopts data gaming to ensure the balanced data traffic. The general step is to establish a joint feature library of network traffic data through the extraction of special data sequences. However, according to the classification, improve the control ability of channel balance, ensure the balanced data channel transmission load, and realize the balanced scheduling of network traffic big data. Although this method can achieve certain effects in the initial stage, as the amount of domestic network data and the type of terminals keep changing, the problem of its small traffic control domain has led to the inefficiency of its data balanced scheduling. In addition, once the data flow exceeds its control mechanism, the problem of big data network load still occurs, for which a new balanced scheduling strategy for network traffic data is urgently needed. The set-pair analysis is an idea of transforming analysis by forming pairs of sets with the association. In this study, a load scheduling regulation method is proposed.
2. Topology Awareness Analysis

Network topology can be regarded as a general term for network big data (one kind of data that includes network information) ponding links, as well as the relationship information of each link node. In general, the three core components are node object, node port and network link path. Through topology awareness analysis, the node situation, link status and data host operation of network data can be determined, which provides a framework basis for data traffic scheduling to reduce network load.

The design adopts Link Layer Discovery Protocol (LLDP) as the core of network topology awareness analysis, introduces data exchange protocols, places topology data and host information data in the exchange unit, and connects the prepared network protocol devices through traffic feedback data. These protocol devices obtain the feedback data packets through LLDP and store the packets centrally.

According to the traditional experience, the use of LLDP to implement network topology analysis probing is to create additional DPID data files and contact packets of the return protocol, together with LLDP packets, they will be put into the Packet data message file, and then connect to the corresponding network devices.

After the network device completes information transmission and reception, it needs to first verify the corresponding data flow and data storage browse table to determine whether there is a feature matching data file. Because the link-layer protocol generally does not set the feature LLDP flow table data, the design needs to be in the open switch device in advance, to ensure that the data packet can resolve the corresponding contact discovery information and protocols according to the DPIO file execution code and file port model. After aggregating the information, message aggregation can be performed through the switching device.

The switch device needs to be added the pin function, the general configuration needs to carry out the GPIO data bus expansion, its direction is the same as the collection input, the function is to accept the link data protocols under all kinds of ports. The switch model is a LOC559 controller with an internal register PPIN for adjusting the switch volume. The execution settings take the data transmission pins of the controller as a collection configuration of basic input functions and then adopts repeated cycles and other ways to collect and output different control information, the execution of the networked control system generally includes program start, system file reading, definition transmission, and quantitative data collection, fair distribution of data, vector control, etc. In general, through the above steps, the core protocol can be completed data loading, thus understanding the entire network topology of the current data network.

In order to ensure the final load balance of the big data of network traffic, the design for the whole network topology information acquisition needs to include network status information, gateway execution, etc., to ensure that the end-user can achieve a balanced distribution of traffic load through the network link status [2].

The design adopts OPENPLOW as the sensing protocol to analyze the network structure, its core is to analyze the port traffic of switch and network device, and to measure the total amount of overall network information through the periodic execution information. Based on this, the network bandwidth is measured, and the difference between the network information data is determined to obtain the bandwidth value, and the difference between these values and the traffic value is the residual broadband [3].

In addition to the residual broadband obtained above, the network load and delay information parameters are also one of the core parameters of the network topology information structure, and the PING data command needs to be used for the detection of the relevant data. The design uses the average value of the data transmission delay between the computer and the switch to obtain the corresponding average value. In order to reduce the measurement requirements, the design adopts the time-stamped subtraction method for the calculation of the average value, and the specific measurement steps are as follows:

The controller sends LLDP data to switch A. The data information includes the corresponding time
information. After switch A gets the time information data, it sends it to switch B. B performs abnormal matching of data information and returns the data to the core controller.

The time of data transmission and data reception is subtracted to obtain the corresponding data difference T1. In the same way, the corresponding opposite time difference T2 can be obtained. The controller and switch A and B will distribute the information, and after the corresponding data reception, the information file will perform the recovery operation. The controller can directly measure the corresponding control information through the time file. In this way, the time difference Ta and Tb between the controller and switch can be obtained. At this time, the expression formula of data transmission between A and B is as follows:

\[ S = T_a + T_b - T_a - T_b \]  

(1)

The average value of current network delay is calculated as:

\[ S' = \frac{T_a + T_b - T_a - T_b}{2} \]  

(2)

3. Data Transmission Calculation under Cloud Computing and Set-pair Analysis Strategy

After obtaining the global network topology structure, a cloud computing algorithm is used for data transmission based on the set-pair analysis strategy. In order to be able to effectively prevent problems such as network loop occlusion, the design adopts a limited loop-free KSP cloud computing algorithm. To ensure the smoothness of the network loops, the design adopts wireless KSP cloud data computing strategy [4]. Set the network nodes as N and let i be the node information data. Now, \( i = 1, 2, 3 \cdots, N \). Let 1 be the starting node and N is the final node. Then there is a relationship formula as follows: (1)

\( \cdot (i) \cdot (j) \) and \( i \neq j \), the relation refers to the path from far point i to j, let \( d_{ij} \) be the relational distance between nodes i and j.

It is certain from the relationship of \( d_{ij} \) that if \( d_{ij} \) is real, it can be expressed as a positive integer, otherwise, \( d_{ij} \) is a value of positive infinity, \( A_i^k \) expresses the value of the deviation of the relationship between network node i and \( A_i^{k-1} \). \( R^k \) is the path root of \( A_i^k \). If \( d_{ij} \) is real, \( d_{ij} \) is expressed as a positive integer, otherwise, \( d_{ij} \) is a value of positive infinity, \( A_i^k \) is the deviation value between all network nodes i and \( A_i^{k-1} \). \( R^k \) is the path root of \( A_i^k \), it can be used to express the same network transmission subpath as \( A_i^k \), that is, the shortest path of each node from 1 to i is same as \( A_i^k \).

If the value wants to be confirmed, \( A_1^k, A_2^k, A_3^k \cdots A_i^{k-1} \) need to be calculated first.

Its calculation formula is:

\[ A_i^{k-1} = \frac{d_{ij}}{e(dij)} \int_0^{di} \frac{e^j N}{A_i + A_j + A_k}d_k \]  

(3)

In the above formula, \( E \) expresses the transmission vector of data; expresses the total data of the standard path;

Let the final result of the path calculation be K, or its number is greater than K, then the calculation is terminated. If the final standard path data is less than K, divide a path into the list, set the list as A, and store the other paths with B. At this time, if A is 1, the calculation process is obtained as follows:

In checking \( A_i^{k-1} \), the subordinate paths of each node, at this time, if the target path overlaps with the target node, the sequence of the other stages remains the same, set the node distance to infinity. In
this study, the DIJKTRA algorithm is applied to find the shortest path between the node \( i \) and the node \( j \), let the subpath be \( R_k^i \), connect \( R_k^i \) and \( A_k \), move \( A_k \) to the storage list [5].

Compare each subpath in the above list B to find the shortest path, if the path can be found in both A and B and the path in table A has exceeded K, then the path calculation needs to mark the minimum path as \( A_k \), then move it from list B into list A, and continue the calculation on the data in list B until the number of data in table A equals K.

At this point, the path calculation method requires the data bytes of the path as the core weight, set the path weight data as 1, K path data calculated by J can be directly queried results, and then define to all types of data, each path data and branch for two parts are as shown in the figure below. The origin of the path query is S, the target set point t and the shortest path under it.

![Diagram of Path Query Result](image)

**Figure 1** Diagram of Path Query Result

### 4. Reconfigure Links Based on Path Calculation Results

#### 4.1 Traffic Routing Control

The integrated control trend of traffic routing is ensured by separating the core key indicators of the link and quantifying the link range, its quantification formula is as follows:

\[
\rho(l) = \frac{\text{AVE}(l)}{C_l} \quad (4)
\]

In the above formula, \( \rho(l) \) expresses the core of current network traffic link \( l \), \( C_i \) expresses the actual comprehensive load rate of the current link data transmission, \( \text{AVE}(l) \) expresses the load expectation of the defined link.

To correctly analyze the overall quality of the network data link, it is possible to prove the smoothness of the transmission terminal of the network traffic data based on the load expectation value, the higher the expectation value, the higher its quality. The design adopts \( \text{AVE}(l) \) as the maximum ratio of the current traffic of each network node.

In addition, this design replans the link weights in order to reduce the traffic load under the premise of using multi-layer link transmission. The traditional network data traffic weights are generally
determined by the link variation value and the inverse of broadband, and the application needs to
determine the relationship between the inverse of broadband and the weights. When the weight is
reduced, the data traffic is enhanced, but after it meets a certain threshold, according to the above core
requirements and definitions of $AVE(l)$, the data link is likely to be overloaded. At this time, the
weight of this link and the inverse of the remaining bandwidth need to be applied as the correction
factors, further increase the weight value, for which the design optimizes its objective function:

$$F(r'_j) = \sum_i \sum_j r'_j \sum_l H'_{ij} \frac{\rho l}{C_i} \quad (5)$$

In the above formula, $r'_j$ expresses that in the lower node $i$ of data network link, the network data
distribution rate on each effective path $j$; $H'_{ij}$ mainly expresses the actual data elements in the current
routing matrix A, its specific meaning indicates the node $i$ under the path of the current network node,
choose the communication link $j$ in the transmission path $j$; the routing matrix A is set with the
specification of $J \times I$ dimensional 0-5 matrix, $J$ expresses the total amount of objective link data in
which network data; $I$ expresses the number of pairs of network data source nodes; if the current
network data links all pass through $l$, it means that in the above formula, the value of $H'_{ij}$ is 1,
otherwise, the average value is 0. $C_i$ expresses the remaining bandwidth value of the network link, its
final optimization formula is:

$$\min \sum_i \sum_j r'_j \sum_l H'_{ij} \frac{\rho l}{C_i}$$

s.t. $\sum_l H'_{ij} \frac{\rho l}{C_i} \leq C_i, \forall l \quad (6)$

In the above formula, $r'_j$ is the target value to be solved; because of the modern network link load,
the carrying capacity of the network cannot be exceeded, so the above constraint concept is proposed.

In the above formula, $r'_j$ expresses the analysis target value. According to the concept of network
constraints, the design adopts the even function analysis formula to solve the above target formula
function, analyze the overall data path packets between the logarithms of nodes, and obtain the
characteristic coefficients of data traffic according to the network controller to indicate the contact data
analysis packets between each path and the switch. The control process steps of the design are as
follows:

step1: Calculate the sending efficiency $r'_j$ for each major source of the current network link $l$ based
on the network link measurement information;
step2: update the current network link cost;
step3: bring the updated link cost value, and calculate the source-end traffic demand cost;
step4: update and calculate the data source end, and assign the rate;

In order to effectively reduce data correlation, the step parameters in the above algorithm have their
independence from each other.
4.2 Link Connection

According to the control policy of network routing, the network traffic weights are centrally assigned to determine the link groups. Since a single weighted link group cannot match the traffic data, the design reconstructs the link connection by designing the link connection work in a static domain name resolution environment. In this regard, the integrated transmission link is created using the status of co-existence of network traffic load information and network information entity information input.

The calculation adopts ARMA data template sample to obtain a sample reconstruction formula for the current information registration data with the DNS data traffic information table as the carrier as follows:

\[
x_n = x(t + n\Delta t) = h\left[z(t0 + n\Delta t)\right] + w_n
\]

In the above formula, \(h\) expresses the effective control of network traffic load, \(w_n\) expresses the overall error that may occur in the process of dynamic domain name resolution. According to the set of feature data mapping, the traffic resolution formula of the current data sample can be obtained as follows:

\[
U = \{U_1, U_2, \cdots, U_N\}
\]

In the above formula, \(U\) expresses the load value of the overall network link input traffic of characteristic dimension \(d\). Based on the above splitting formula, the network channel structure is constructed as follows:

![Diagram of Channel Structure](image)

According to the above channel structure, the constraint of feature traffic is constructed by minimizing the input and output data as delay along. Then the filter matching strategy is adopted to measure the total amount of filter according to the network load, the results of the overall extremes are collected, and the set of load filter features of the network traffic is specified, and its set expression formula is:

\[
x_n = a_0 + \sum_{i=1}^{AR} a_i x_{n-i} + \sum_{j=0}^{AR} b_j \eta_{n-j}
\]

In the above formula, \(a_0\) is the maximum tolerance value of network channel load; \(x_{n-i}\) is the traffic measurement scale; \(b_j\) is the traffic sequence value. By the above formula, the network link is reconstructed.

The domain frequency expansion is adopted to perform the balanced load frequency domain expansion under DNS domain name, its expansion formula is:

\[
f(x) = \sum_{i=1}^{l} (a_i + b_i) k(x - x_i) + b
\]

\(k(x - x_i)\) is the maximum value of network load pulse, \(b\) is the traffic disturbance; \(a_i\) is the data analysis value; the load characteristics of the current network traffic are decomposed as the following formula:
\begin{align}
\int_{-\infty}^{+\infty} W_{i} (t, v) dt = |X(v)|^2 \\
\int_{-\infty}^{+\infty} W_{j} (t, v) dt = |x(v)|^2
\end{align}

In the above formula, $|X(v)|$ is the predicted value of load balance. According to the analysis formula, the network traffic link can be reconstructed, and the relevant data traffic control method can be further adopted to achieve the balanced scheduling of big data of network traffic.

5. Conclusion
This research performs a study on the load problem of traditional big data network transmission link and proposes a new balanced load strategy. Its core lies in the calculation of the network topology and the analysis of link weights to propose the optimal solution and finally improve the reorganized links. It is proved that the method can effectively improve the load problem of network data transmission. In the future, the weight matching degree and the data transmission efficiency need to be further improved to achieve continuous optimization.

References
[1] Wang Wenting, Jing Junshuang and Zhang Hao. A Balanced Scheduling Method for Network Traffic with Large Data Based on Set-pair Analysis [J]. Automation and Instrumentation (1):4.
[2] Liu Yan. Simulation of Network Scheduling Algorithm Based on Traffic Tilt Classification [J]. Computer Simulation, 2013, 30(11):289-292.
[3] Cao Shaohua, Lu Qinghua and Zhang Hongxia, et al. SDN-based Dynamic Traffic Scheduling Method for Server Clusters [J]. Journal of China Academy of Electronics and Information Technology, 2016, 11(006):625-628.
[4] Dong Qian, Li Jun and Ma Yuxiang. Centralized Control-based Traffic Scheduling Method for Named Data Networking [J]. Journal on Communications, 2018, v.39; No.373 (07):72-84.
[5] Xu Lei, Li Qing and Xiao Xiangze. Research on Big Data Policy Co-word Networks Based on Grounded Theory [J]. Journal of Modern Information, 2018, v.38; No.324 (06):159-166.