Datacenter network based on Micro-Electro-Mechanical System optical switching

Weicheng Jiang[a], Lanying Li[b], Yong Li[c], and Weishu Guo[d]

The Engineering & Technical College of Chengdu University of Technology, Leshan Sichuan 614000, China

[a]leshanlv@163.com, [b]lly_16@163.com, [c]994087842@qq.com, [d]305358837@qq.com, email

Keywords: Datacenter network, Optical circuit switching, Micro-Electro-Mechanical system

Abstract. The size of datacenter is growing rapidly. Datacenter network faces enormous challenges in the communication capabilities. Micro-Electro-Mechanical System (MEMS) optical switching technology is an idea for the network. In this paper, MEMS optical switch is used to construct the datacenter network. The variant B tree is used in the access to balance the traffic. The hierarchical processing technology is adopted to reduce the complexity and improve efficiency. The structure and control mode of the network are described in detail. The model can adapt to the network environments, especially in the large flow, which can improve the efficiency of optical switching. The simulation tools are used to test the network. The results show that the network has good performance.

Introduction

Datacenter is the infrastructure of modern society, which has an important role in people's production and life. The new service is emerging and people's demands continue to improve. The size of the datacenter is developing very fast. Some datacenters run tens of thousands of servers. This brings great pressure to the communication of the datacenter network, which requires high bandwidth and high speed of the network. There are many disadvantages such as low communication bandwidth and limited capacity in the electrical switching technology. Optical switching technology has high switching capacity and switching speed, and less energy consumption. Now, optical switching technology has gradually entered the practical application, MEMS optical switch has been used in industrial environment.

Some scholars have studied the optical switching network. The fundamental requirements of the hybrid architecture and their design options were discussed in paper[1]. The flow control system was used to detect the flow between the cabinets and the traffic demand matrix was constructed. A hybrid electrical/optical switch architecture was introduced in paper[2], which used Edomnds algorithm to calculate the optimal optical link. MEMS optical switch was used for optical lines between pods. Wavelength division multiplex technology was used to increase the bandwidth of optical lines. Arrayed-waveguide grating router (AWGR) optical router was used to construct a domain of all-optical packet switching network in paper[3]. The central control plane adjusted the output wavelength of tunable wavelength converters (TWC) according to the destination address of the packet. The data was transmitted to the corresponding port by the AWGR optical router. A hybrid optical network was proposed in paper[4], which was based on AWGR and TWC. According to the size of the SOCKET buffer when the TCP connection was established, the network traffic demand was detected and the TWC was used to establish the continuous optical circuit between the cabinet.

System Design

MEMS optical switch is an optical device controlled by a micro mechanical device. There are a N×N mirror array in the optical switch. They are attached to the micro motor controlled by an embedded processor, which controls the rotation or deflection of the micro mirror by electrostatic (or drum) force. The light of the input port is redirected to the output port to realize the routing of optical signals. The optical switching device based on MEMS micro mirror has the advantages of low loss, low crosstalk,

low polarization sensitivity, high switching speed, small size and large scale integration. Optical channel adapter (OCA) can carry out the conversion of optical signals and electrical signals for the transmitter and the receiver. Each OCA has a 1:N splitter, which can separate the mixed beam into different wavelengths of light. There is also a receiver array to convert optical signals into electrical signals. In this paper, these optical devices are used to establish the optical switch link, and the data transmission is carried out under the control of circuit switch manager. A unified control switching network (UCSN) is proposed.

MEMS optical switch is a micro mechanical control system. It takes a certain amount of time, usually in milliseconds. If it is too frequent, the consumption of time will be much. By reasonable control strategy, time can be reduced and the efficiency of optical switching can be greatly improved. Aiming at the characteristics of large flow and long duration in datacenter, the control of MEMS is unified in this paper. According to the different state, the system is divided into two phase. One is the MEMS control phase, which time is t_c. The other is the data transmission phase, which time is t_s. In the MEMS control phase, the system is unified to adjust to complete the change from the input port to the output port through the control of mechanical devices. All the input ports in the system are adjusted and changed, as shown in Fig.1. Different input ports are transmitted to different output ports. For an input port, all output ports is transmitted in a cycle. Data is transmitted from the input port to the output port in the data transmission phase.

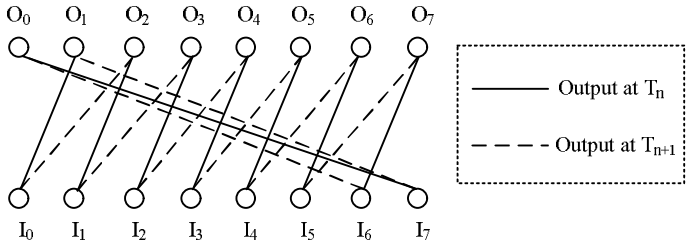

Fig.1. An example of port switching

The output port will be changed to different ports for an input port after time $T = t_c + t_s$. Each input port can be achieved to the data transmission of N output ports after the time NT in N ×N MEMS optical switch. The data of the same input port is transmitted to the same output port in the same period. For the convenience of processing, each input port is provided with N queues and the data transmitted to the same output port is buffered in a queue. The queue number is Q_0, Q_1, Q_2,… Q_{n-1}. The relationship between the output port and the queue number is $O_i = F(Q_i)$.

The data in queue Q_i is transmitted in T_i (t_i<N). The transmission of data is periodic. The relationship between the number of queues and the time (t_i<N and t_i>N) is $Q_i = T_i \bmod N$.

The scheduler works according to the time. Packets in the corresponding queue are sent to the input port. Relationship between the output port number and the time is $O_i = F(T_i \bmod N)$. Packets are able to be routed to the destination port according to the time slice.

Each node has a routing table. The packets are pre-processed. When packets arrive, they are buffered in the corresponding queue according to the export address in the routing table. Packets with the same export address are buffered in the same queue. After the classification of storage, the time to find the table is omitted in the forwarding process and the complexity of processing is reduced.

In order to speed up the efficiency of data transmission, packet bursting is used in the system. A number of packets will be linked to and put together to string out. The packets are assembled together in the MEMS control period (i.e., t_c), which will be sent during the period of t_s. The fill characters are used among packets, so that the receiver can distinguish different packets. It can reduce the waiting time after sending a packet and more packets can be sent in one time slice.

The packets in t_s should be processed as many as possible. Due to the different length of packet, it is difficult to ensure that the packet in the t_s is exactly an integer packet. It is possible to add a packet that cannot be completed within the t_s. It is time to have surplus without adding a packet. Set the length of data to be transmitted in a time slice is L_S and the length of packets assembled together is L_1. It has

$L_1<L_S$. Add a packet P_C. The length of P_C is L_2 and $L_1+L_2>L_S$. There are two ways to process the packet Pc. The first way is that packet Pc is divided into two parts. There are $L_2=L_{21}+L_{22}$ and $L_1+L_{21}=L_S$. The first part of the packet Pc is encapsulated in a time slice, and the rest is assembled and sent in the next cycle. It requires the split and reorganization of the packet. It is complicated. The second way is that the packet Pc is not assembled. Split packets will be sent in the next cycle. In this way, the utilization of time slice is not enough, but the processing is simple. The second way is adopted in this paper.

To balance the traffic of input ports, the accesses are organized into the variant B tree structure. The root of the variant B tree is connected to a port of the MEMS optical switch. Leaf node is an access point. There are several reasons for the variant B tree structure. First, the number of ports in the switch is within a certain range, which corresponds to the subtree being limited. Second, most switch ports should be attached. Intermediate nodes have at least a certain amount of subtrees. Third, if part of the branch is too long, delay will be large. The depth of each access layer should be consistent, which corresponds to the leaves of the B tree appearing in the same layer. Therefore, this kind of structure can meet the requirements of practical application. The height of the tree is restricted, which avoid that the "vertical level" is too deep. The B tree is modified in the paper, which is called "the variant B tree". The requirements are as follows. (1) Each node in the tree has at most M subtrees. (2) All intermediate nodes have at least $\lceil m/2 \rceil$ subtrees. (3) All leaf nodes appear at the same level. (4) The height of the tree is H, which is H<h (h is the threshold). (5) With the intermediate nodes of the same layer, the flow difference of each node is less than f (f is the threshold). The requirement of condition (5) is helpful to keep the total flow difference of each node is very small, which is good for load balancing. Set R_i as the traffic of subtree i. The total flow of a switching node is $U_j = \sum_{i \in k} R_i$. K is the number of subtrees. If intermediate nodes j_1 and j_2 are the same layer, there is $\left|U_{j_1} - U_{j_2}\right| < f$ .That is any two intermediate nodes of the data flow within a certain range of f. Fig. 2 is a variant B tree. The variant B tree has three layers and the largest subtree node has 4 children and the minimal subtree node has 3 children.

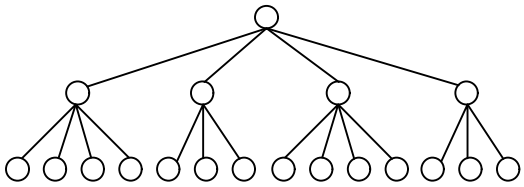

Fig.2. An example of the variant B tree

In order to evaluate the data traffic, the traffic monitoring is carried out at the access node. The switching node contains a group table and one or more flow tables. The switching node receives packet and query flow table in sequence. If a match is successful to a list of items, the corresponding statistics for the table entry are updated, including the number of packets and the length of the data. After the corresponding operation, packets are forwarded. The switching node counts the total number of packets, the number of bytes and the number of packets dropped in a period of time. Traffic information can be used to organize the variant B tree.

Data arrival is a Poisson distribution of parameters l and the service strength of m in the system. By queuing theory, the average buffer length of the queue is $L_q = \frac{l^2}{m(m-l)} \cdot \frac{1}{N}$ and the average waiting time for data is $T_q = \frac{l}{m(m-l)} \cdot \frac{1}{N}$. N is the number of queues.

Evaluation

In this paper, Opnet and Matlab are used to evaluate the performance of the network. To simulate the diversity of network flow, the arrival time interval of flows are distributed by Constant and Poisson, and each distribution is provided with a plurality of parameters to generate data streams. ESM is the result of the electrical switching mode, and UCSN is the result of the design of the network in the paper.

(1) Overflow

The overflow shows the relationship between service capability and service request. Overflow is large. Many service requests are not responded and many packets are discarded at the switching node. The capacity of the switching node is insufficient, which affects the performance of the network. The result is shown in Fig.3. The overflow of UCSN is less than that of ESN. The speed of the switch in the electrical field is lower than that of the optical field. The ability to switch is insufficient in ESN. When the service request increases, many packets will not be able to be responded. Overflow is more than that of UCSN. In addition, a unified control of the MEMS control is used in the UCSN, which increases the capacity of data transmission. The switching nodes are organized into the variant B tree structure in the access, which the traffic is balanced. These measures can effectively reduce overflow. Therefore, the overflow in UCSN is significantly less than in ESM.

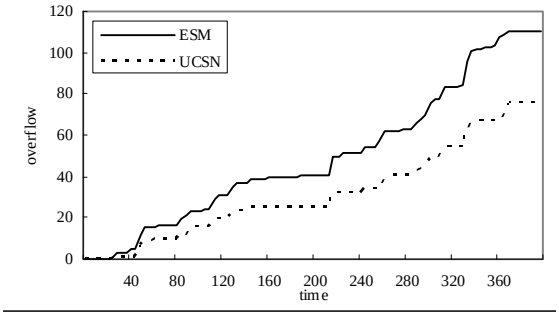

Fig.3. Overflow

(2)Delay

The data is processed at the switching node, which brings the processing delay. If delay is small, the switching speed is high and network performance is good. The average delay is shown in Fig. 4. The average delay in UCSN is much smaller than in ESM. The speed of optical switching is faster than that of electrical switching. The corresponding measures taken in this paper also help to reduce the delay. Thus, the average delay in UCSN is much smaller than in ESM.

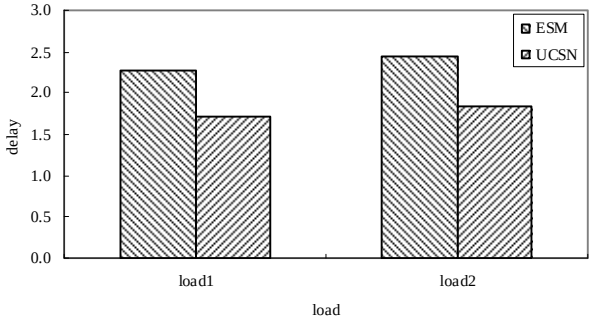

Fig.4. Delay

Conclusions

In view of the characteristics of large traffic in datacenter, MEMS optical switch is used and a datacenter network model is constructed in the paper. The unified control of MEMS devices is achieved and the complexity of processing is reduced. It is helpful to improve the efficiency of the system and the speed of the network by using the hierarchical processing and pretreatment technology. More traffic can be transmitted. In order to assist the MEMS optical switch, the variant B tree structure

is adopted in the switch access. It is conducive to balance the network traffic and bring good performance and expandability for the network. The simulation tool is used to evaluate the network model. The results show that compared with the traditional structure of the network, the proposed scheme can reduce the average network delay, reduce the packet loss and improve the throughput of the network.

Acknowledgements

This work was financially supported by the Education Department of Sichuan Province (16ZB0412) and the Engineering & Technical College of Chengdu University of Technology (C122015007).

References

[1] G. Wang, D. G. Andersen, M. Kaminsky, et al. Acm Sigcomm Computer Communication Review Vol. 41(2011), p. 327

[2] N. Farrington, G. Porter, S. Radhakrishnan, et al. Acm Sigcomm Computer Communication Review Vol. 41(2010) , p. 339

[3] X. Ye, Y. Yin, S. J. B. Yoo, et al. DOS: a scalable optical switch for datacenters, ACM/IEEE Symposium on Architecture for NETWORKING and Communications Systems, ANCS (2010)

[4] Dawei Zang, Zheng Cao, Zhan Wang, et al. Chinese Journal of Computers Vol. 39(2016) p. 1868 (in Chinese)