A Dynamic Multipath Scheduling Protocol (DMSP) for Full Performance Isolation of Links in Software Defined Networking (SDN)

Syed Asad Hussain, Shuja Akbar, Imran Raza
Department of Computer Science, COMSATS Institute of Information Technology Lahore, Pakistan
{asadhussain, shujaakbar, iraza}@ciitlahore.edu.pk

Abstract— **Software Defined Networking (SDN) has emerged to be an ultimate solution for the management of data centers. The separation of control plane from the data plane has made it possible to manage physical and virtual networks through the SDN controller. This paper presents a novel Dynamic Multipath Scheduling Protocol (DMSP) for effective scheduling of packets on SDN virtual links, thus achieving full performance isolation. The proposed solution reacts and adopts appropriate scheduling strategy during transitions. The source/destination mapping for path selection ensures service specific bandwidth for each pair of hosts in a multipath environment. DMSP makes real-time changes to select a least congested link for routing of data packets. The simulation results show that DMSP has better average end-to-end delay, packet drop ratio per flow, packet delivery ratio and packet drop ratio as compared to other scheduling techniques thus it improves resources utilization.**

Keywords— SDN; full performance isolation; multipath routing; data center networks.

I. Introduction

Management of the virtual links turns out to be a time taking task for network managers. SDN [1] has emerged to be an effective solution for bringing automation in management context. SDN approach has been ensuring management of the network resources without administration intervention. It has reduced the resource management complexity while retaining its ability to model physical network as a software entity. Separation of control plane from the data plane resulted in real time management of the network according to changing network conditions. At the same time load balancing also looked after traffic engineering in real-time with varying load demands. Scheduling and load balancing techniques are used to improve network performance. SDN needs dynamic scheduling strategies to optimize network performance.

Multiple paths of equal cost are available in a typical data center topology for packet forwarding. Multipath scheduling in SDN is more challenging as compared to conventional networks, since there is no default module in the SDN controller using multiple paths. Using single path forwarding approach in the presence of redundant links for same destination leads to underutilization of the network resources. There are various approaches such as Hedera [2], Mahout [3] and ECMP [4] to utilize redundant paths. A standard approach is ECMP [5], but it comes with some potential problems discussed in RFC 2991 [6] including variable latencies, variable path Maximum transmission Unit (MTU), and debugging. The solutions proposed so far use hash based decision making for selection of the paths. The problem with such approach is that multiple large flows can be hashed to same link and may overload a path in the network. The challenges remain to be addressed are overcoming the issues of traditional ECMP [5] approach and mapping of multiple large flows onto the same path dynamically.

The proposed Dynamic Multipath Scheduling Protocol (DMSP uses bandwidth aware strategy to prevent mapping of the flows onto the congested links. DMSP implements a central controller that keeps track of all congested links and links' delays. The emerging flows can be redirected or mapped to the least congested links in a data center topology. DMSP makes an effective use of the network resources and improves quality of services. It is a reactive approach that makes managing of the links easy in a data center topology.

- Full performance isolation ensures maximum delivery of the packets in data centers.
- Reactive approach dynamically alters forwarding rules for the flows emerging for a longer time in the network.
- Source/Destination port mapping and round-robin strategy provides optimal mechanism for the usage of the network resources.

This paper is divided into following sections. First section is introduction of the paper, section 2 discusses related work. In section 3 proposed solution is presented. Section 4 is based on simulation results and section 5 concludes the paper.

II. Related Work

In [2] Hedera is presented as a dynamic scheduling algorithm and adaptively schedules multi-stage switching fabric in order to efficiently use network resources. Hedera uses available multiple paths of equal cost to implement dynamic scheduling by collecting flow information from the swi tches and instructing them to re-route traffic accordingly. It detects large flows at edge switches, when a new flow is initiated the switch forwards it along one of its equal cost paths.

In [4] an analysis of Hash Based Equal Cost Multipath forwarding solution in SDN is performed. ECMP is a technique of splitting flows across available paths using flow hashing.

978-1-5090-5541-8/17/$31.00 ©2017 IEEE

Hash based scheme enable switches to have multiple possible forwarding paths available for a given subnet. ECMP was the first prominent solution proposed to utilize available multiple paths for load balancing. ECMP was a static load balancing approach that did not consider traffic or load on the switches. Hedera [2] and Mahout [3] were proposed later and these load balancing schemes considered traffic analysis to deal only with the elephant flows to avoid controller's extensive interventions.

In [7] a scalable commodity data center network architecture is presented. Data center networks contain several connected computers and significant aggregate bandwidth. The proposed solution achieves full bi-section bandwidth as interconnecting switches in a fat-tree (data center topology) architecture. The redundant paths may also be non-blocking, non-blocking here refers to arbitrary communication patterns, there are some sets of paths that soak all the bandwidth available to the hosts in the network topology.

In [8] load balancing based on Open Flow protocol for the fat tree topology with multipath support is proposed. Data center networks are designed to meet demands of the densely interconnected hosts. Network topology and routing algorithms have ultimate effect on the performance of data center networks. Fat tree topologies are the widely adopted structures for the data centers and the proposed strategy is the way to achieve high performance and low latency in the network. Load balancing in fat tree topology can never be fully achieved using traditional approaches.

In [9] an adaptive multipath routing algorithm based on collaborative load balancing and topological characterization is presented. A multipath routing algorithm finds multiple equal cost paths between two end nodes in the network regardless of the constraints on the cost of these multiple available paths.

In [10] a loss free solution for data center based on multipath approach for SDNs is proposed. In the proposed solution link layer, multiple-paths are introduced over a spanning tree. Such a solution has emerged to be viable replacement for data center networks. A two-tier solution is proposed that integrates multipath based load with congestion control. Path load and buffer level are used to trigger multiple paths and congestion control instead of using two separate parameters for the purpose of load balancing.

In [11] Depth-First Worst-Fit Search based on multipath routing for data center networks is proposed. DFS+worst-fit-link selects a path using the DFS algorithm that may lead to unnecessary backtracking which increases computation time. The proposed algorithm uses greedy approach and takes only local view into the account to select a path. The selected path may not utilize highest available bandwidth.

In [12] SDN-based adaptive worst-fit multipath routing (AWMR) algorithm is proposed which selects multiple paths for managing flows in data centers. It works in two stages; in first stage, it finds the shortest paths and computes available bandwidth while in 2nd stage the algorithm utilizes a widest disjoint path (iwdp) scheme to select worst-fit paths. However, the algorithm may underperform due to variable path MTU, variable latencies and buffering requirements.

III. Multipath Routing protocol for Full Performance Isolation of Links in Data Center Networks

Full performance isolation in multipath routing of links in data center networks is implemented to resolve the issues of traditional approaches. The solution works as a multipath scheduling algorithm that utilize multiple paths to deal with the huge amount of traffic in data centers. The SDN controller keeps record of link utilization, end-to-end delay and other links characteristics. It prompt changes in packet scheduling by least congested links to serve the flows. The RESTful API [13] of SDN controller makes it easier to implement such a dynamic solution for managing the links optimized for flows to be handled distinctly. A fat tree data center topology is used to bring redundant links between two nodes in the data center topology. Fat tree network topology can be segmented into the three layers.

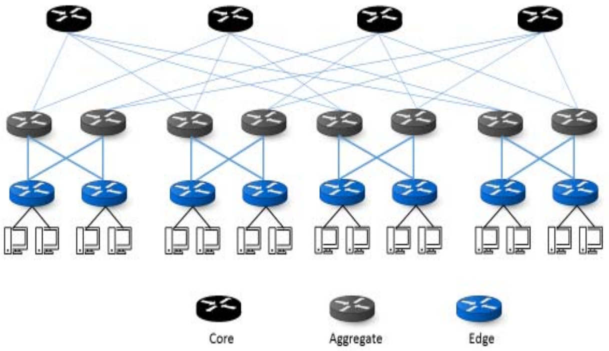

Fig. 1. Data Center Topology

Three layers of switches are introduced to provide redundant links between a pair of hosts. To overcome the issues of splitting load into multiple links in traditional approaches, source port/destination port based mapping is used as a solution for unicast forwarding. Such a technique introduces performance isolation for each service to introduce each flow with a distinct path that will avoid the ordering issues at receiver end and resolves problem of multipath approach (single input/multi-output). The standard applications SDN control plane are on Layer 2 forwarding (data plane) where MAC address of the hosts are used to map source and destinations. The solution works as IP based routing that overcomes the problems of handling flows on each hop. The proposed solution works on the principle of round-robin strategy to eliminate the problem of mapping flows onto a single path generated by different clients. It provides a dedicated best available path for each client to handle traffic to destination. Packet forwarding on the basis of port mapping gives each application with a distinct path to transfer data.

DMSP is implemented as a SDN application and it is emulated using Mininet [14] that provides real network environment on a single machine. The SDN application is written as software program for SDN controller that constructs routing tables and instructs the data plane where packet should be forwarded. It works in both proactive and reactive modes

(Hybrid Scheduling algorithm) to shape the network traffic. In proactive mode, routing entries will be recorded once the topology is created and the hosts in the network are detected. Once a new flow is detected, SDN controller application will detect shortest paths and update the routing table. The proposed solution will not only provide high throughput as compared to single path, but also resolve the issues of traditional multipath approaches in SDNs.

```
(a,b) = Pair of source host a and destination host b
H= set of hosts exist in topology
P(a,b)= Set of shortest path between host a and b

W(e) = load on the link (path)

F = Flow Table

T = fixed time interval

BEGIN
Discover all nodes in topology where hƐH;
FOR EACH h
        Calculate shortest path for each pair of node;
        Put path into P(a,b);
        F = CreateFlowTable (P(s,d));
        Map source and destination IP fields;
        Map source port and destination port;
FOR EACH flow f
        Lookup(F);
        // is flow entry is available
        IF seen then;
        Send packets from source using path p+1 Mod $\Sigma$ ($P_i$);
        // shortest equal cost paths
        ELSE
        Generate New Flow
        End IF
After each interval T
        Calculate W(e) for each path
        Update flow table
END
```

Objective of DMSP is maximization of resource utilization in data center networks. The solution is implemented as a SDN controller application. Floodlight SDN controller [13] is used as controlling entity in the proposed solution, and the multipath solution is implemented as part of Floodlight controller that acts as a brain in SDN environment. The objective is to take advantage of the separated control plane and data plane in software defined networking approach. The solution is implemented as a part of Floodlight controller that will schedule and route traffic in the network for optimal use of resources with minimal losses.

The SDN application (Floodlight module) creates routing tables based on the global view of the network topology (Fat Tree Topology). The appropriate routing table is installed in each SDN switch. Each switch forwards packets according to the routing table installed by the application. In other words, flow rules are pushed onto each switch for the forwarding rules.

Each entry in a flow table has match criteria (on the basis of IP and other headers) which defines rules to be applied to the packets. Generally, each entry also has one or more instruction/action which is/are applied to each packet that matches a rule. In a module entries are installed that match packets on the basis of their IP address and the Ethernet type. The output action is based on sending the packet to a specific port on the destination host.

The selection of the path is made on the basis of round-robin policies after calculating all shortest paths from the source to a destination. The mapping of active connections is performed and identified by the client's IP and port. Once a path has been assigned to a particular host, all packets for that connection should be directed to that host at a specific port.

The controller module's statistics are used to identify congested links in data center. The flow table is updated accordingly to bring least congested links before highly congested links. Controller examines each link for the transmitted bytes against capacity of the link to measure the load of a link. The continuous monitoring of links may bottleneck controller, so to ensure scalability, statistics are collected and flow tables are updated at fixed intervals.

Floodlight works in both proactive and reactive modes for the scheduling of data packets. It works in a proactive mode, once the topology is created, SDN controller finds shortest paths for all the attachment points for the source and destination devices. The shortest paths between each pair of hosts (source/destination) are calculated and the flow rules are installed on each intermediate switch including edge, aggregate and core switches.

Whenever, *PacketIn* class of floodlight controller is received with an unknown destination, the packet is flooded and appropriate rules will be installed on the switches. The baseline scheduler works on the round-robin strategy, instead of using a single shortest path to deal with all flows from all devices from the same island (core, aggregate and edge layers of switches). The proposed strategy works on the basis of flows and it deals each flow separately and distinctly.

SDN controller is used to provide flows with least congested paths. The solution improves big data performance as the least congested path not only ensures maximum bandwidth but also overcomes the issues of ECMP approach. However, tiny flows may face slight delays due to updates in flow table. All the dynamic approaches make tradeoff for better management in data center environment, where large flows are under consideration.

IV. Simulation and Results

The proposed solution is tested with the help of emulations using the Mininet tool. A Fat Tree topology consisting of 4 pods as shown in Figure 1 is used to create a data center network. Topology has been created in Mininet using python code and CLI is used to run the emulation. Network topology can be divided into the 3 layers. In first layer edge switches are connected with hosts (two hosts connected with each Edge Switch). In the upper layer, there exists aggregate switches that are used to introduce alternative paths. At the highest layer of the topology there are core switches. Network topology is created using the command line while it bypassed the default controller of Mininet and used Floodlight as remote controller. Mininet CLI Tool is provided with the remote IP address and port number to run controller remotely which control the network.

Floodlight provides a web interface that is used to get statistics, check topology and flow information as the output of environment. IPERF is used to be a tool for generating traffic in the simulation. There are various features, and constraints that can imposed on the network topology using the IPERF tool. IPERF is used for measuring TCP and UDP bandwidth performance. The traffic engineering in the data center topology is performed via XTERM. XTERM program is a terminal emulator for the X Window System. It facilitated us to generate traffic from the hosts.

The proposed solution is compared with traditional single path and multipath algorithms such as DFS Worst Case approaches [11] and AWMR (Adaptive Worst-fit Multipath Routing) [12] with same parameters and constants.

A. Average End-to-End Delay

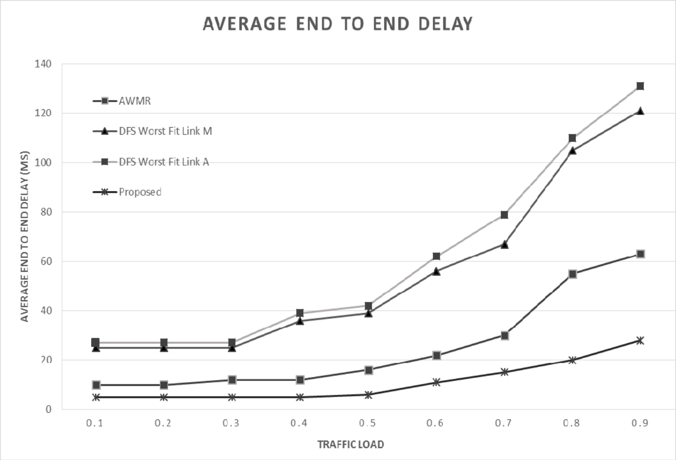

Fig. 2. Average end-to-end delay

Figure 2 shows average end-to-end delay of the network for all the approaches. It depicts the behavior of the network from less to highly congested network environment. The comparison is with three other multipath techniques. The proposed solution has worked better because of the intervention of the controller that introduce least congested links to serve flows in the data center networks. Ultimately full performance isolation as proposed ensures highest bandwidth utilization hence the improved throughput of the network.

B. Average Packet Drop Ratio Per Flow

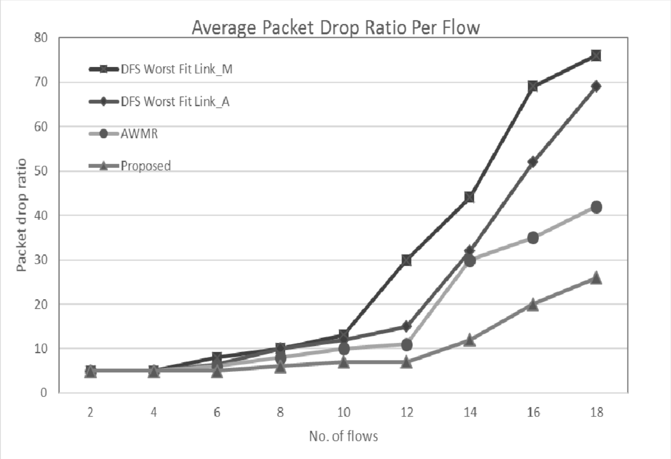

Fig. 3. Average Packet Drop Ratio per flow

In Figure 3 average packet drop ratio per flow is least for the proposed scheme as compared to Depth-First Worst-Fit Search based multipath routing approaches [11], and AWMR [12]. The proposed solution ensures maximum usage of resources with the increase in the traffic. The proposed approach reduces the average packet drop ratio per flow and the full performance isolation of the links ensures maximum bandwidth for the flows to utilize.

C. Packet Delivery Ratio

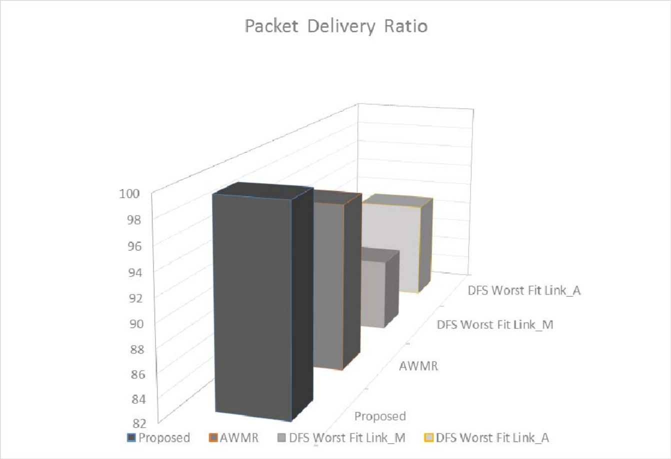

Fig. 4. Packet Delivery Ratio

The packet delivery ratio of the proposed solution is recorded up to 99% in the data center network. The higher the packet delivery ratio, the lower will be the packet drop ratio hence less re-transmissions in the network.

D. Packet Drop Ratio

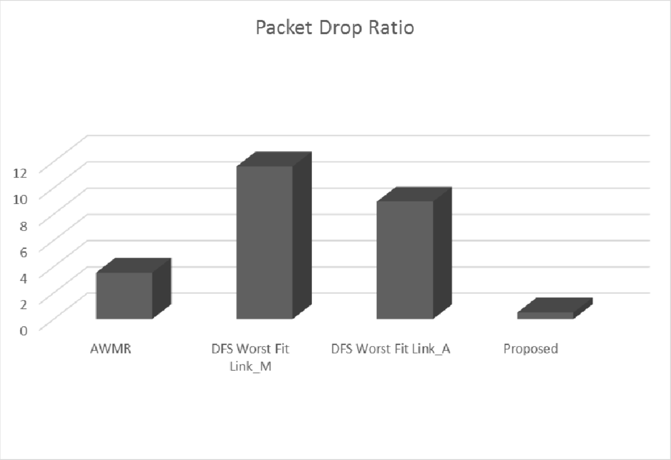

Fig. 5. Packet Drop Ratio

The proposed solution ensures maximum packets delivery and drop ratio of the packets in the proposed solution is less than 1%. The reason behind the least pocket drop ratio is the allocation of the dedicated path and not link for each flow that ensure delivery of the packets at better percentage. To avoid packet loss in the data center network (Fat Tree Topology) the packets flow is sent to the path with maximum available bandwidth.

V. CONCLUSION

A novel scheduling algorithm is implemented as a controller application in SDN environment. Unlike the traditional approaches, proposed algorithm works effectively for the management of the load on the links via IP based routing and path allocation based on the source port/destination port mapping. In data centers, a topology is used that features redundant links that should be utilized effectively for managing high volume of traffic. A common problem in the single path adoption is limited use of the resources, such a strategy may not only increase the cost, but also brings some potential challenges in managing the traffic. Although, control plane and data planes are separated in SDN but a lot of intervention such as by Hedera creates bottlenecks for the controller and links. Sampling could be a good technique for demand estimation and load measurement on the links. The approach brings each pair of hosts with at least one dedicated path until number of pairs becomes equal to the number of alternative paths. It works as a hybrid solution for the load management on the links. It is static in operation that reduces controller's intervention by working in proactive mode in the start when a topology is created. The solution also works in reactive mode, once a flow that is not seen before and there is no routing entry in the routing table, in such a situation controller plays a role to update flow rules for the newly emerged flow. The overall performance of the data center network is improved with the implementation of SDN controller application which is tested via emulating environment as well.

REFERENCES

[1] https://www.opennetworking.org/sdn-resources/sdn-definition (Accessed on September 21,2016)

[2] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, A. Vahdat, "Hedera: Dynamic Flow Scheduling for Data Center Networks". In NSDI'10, Proceedings of the 7th USENIX conference on Neworked Systems Design and Implementation, 2010, pp. 19-19.

[3] A.R. Curtis, W. Kim, P. Yalagandula, "Mahout: Low-overhead datacenter traffic management using end-host-based elephant detection." InINFOCOM, 2011 Proceedings IEEE 2011 Apr 10 (pp. 1629-1637). IEEE. , "Mahout: Low-overhead datacenter traffic management using end-host-based elephant detection". In Proceedings of IEEE INFOCOM, 2011 pp. 1629-1637.

[4] Hopps, Christian , "Analysis of an equal-cost multi-path algorithm", RFC 2992, November, 2000.

[5] Shang, Zhihao, W. Chen, Q. Ma, and B. Wu. , "Design and implementation of server cluster dynamic load balancing based on OpenFlow" In Proceedings of International Joint Conference on Awareness Science and Technology and Ubi-Media Computing (iCAST-UMEDIA), 2013, pp. 691-697.

[6] Hopps, Christian E., and D. Thaler, "Multipath Issues in Unicast and Multicast Next-Hop Selection" RFC 2991, November 2000.

[7] Al-Fares, Mohammad, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture." In Proceedings of ACM SIGCOMM '08, Conference on data communication, 2008, pp 63-74.

[8] Li, Yu, and D. Pan , "OpenFlow based load balancing for Fat-Tree networks with multipath support". In Proceedings of 12th IEEE International Conference on Communications (ICC'13), Budapest, Hungary, June 2013, pp. 1-5.

[9] THEOLEYRE, Fabrice, S. CATELOIN, and P. MERINDOL, "Adaptive Multipath Routing: Collaborative Load Balancing and Topological Characterization". Network Research Group, University of Strasbourg, France, 2014.

[10] Fang, Shuo, Y. Yu, Ch. H. Foh, and K. M. M. Aung. , "A loss-free multipathing solution for data center network using soft using software-defined networking approach." In APMRC, 2012 Digest, pp. 1-8. IEEE, 2012.

[11] Cheocherngngarn, Tosmate, H. Jin, J. Andrian, D. Pan, and J.Liu,"Depth-First Worst-Fit Search based multipath routing for data center networks." In Global Communications Conference (GLOBECOM), 2012 IEEE (pp. 2821-2826). IEEE.

[12] Lei, Yi-Chih, K. Wang, and Y.H Hsu. , "Multipath Routing in SDN-based Data Center Networks", In Proceedings of European Conference on Networks and Communicaions (EuCNC), 2015, pp. 365-369.

[13] http://www.projectfloodlight.org/floodlight/ (Accessed on OCT 21,2016)

[14] http://www.mininet.org. (Accessed on OCT 21,2016)