Leveraging Predictive Analytics to Achieve Knowledge-Defined Orchestration in a Hybrid Optical/Electrical DC Network: Collaborative Forecasting and Decision Making

Wei Lu, Lipei Liang, Bingxin Kong, Baojia Li, Zuqing Zhu

University of Science and Technology of China, Hefei, Anhui 230027, China, Email: zqzhu@ieee.org

Abstract: We design and experimentally demonstrate a hybrid optical/electrical DC network that achieves knowledge-defined orchestration with two collaborative machine learning modules.

OCIS codes: (060.1155) All-optical networks; (060.4251) Networks, assignment and routing algorithms.

1. Introduction

With the fast development of cloud computing, data-center networks (DCNs) are facing exponential increase of computing tasks [1]. To provision these tasks, one needs to allocate not only IT resources in servers but also bandwidth on intra-DC links. This suggests that an effective network/IT resource orchestration mechanism is desirable. Meanwhile, it is known that the traditional DCNs that are purely based on electrical packet switching will be unsustainable soon [2]. To address this, researchers designed various hybrid optical/electrical (H-O/E) DCNs [1, 3]. Specifically, in such an H-O/E DCN (*e.g.*, in Fig. 1(a)), the inter-rack interconnections combine the original electrical packet network with an optical circuit switching network, to explore the advantages of both. Note that, task provisioning can cause highly dynamic traffic in a DCN [4]. For instance, in terms of data-transfer distance, the ratio between inter- and intra-rack traffic could change over time, while in terms of data-rate and life-time, mice and elephant flows would be time-variant too. Apparently, we cannot realize effective network/IT resource orchestration in an H-O/E DCN without the precise knowledge on its traffic characteristics. Moreover, considering the latencies of optical switch reconfiguration and virtual machine (VM) migration, our orchestration mechanism should try to minimize their frequencies and find the best time to invoke them if necessary. This, however, means that only knowing the current traffic characteristics is not good enough, and we need to predict future traffic in the H-O/E DCN precisely. Then, based on the traffic prediction, an intelligent decision making mechanism needs to be designed to achieve effective orchestration.

In this work, we follow the principle of predictive analytics in human brain to design and implement an H-O/E DCN system that can realize knowledge-defined network/IT resource orchestration for task provisioning. Specifically, the proposed system leverages two machine learning (ML) modules and makes them work collaboratively to first predict future traffic (*i.e.*, forecasting based on memory) and then determine the optimal network configuration (*i.e.*, decision making based on knowledge). The H-O/E DCN is based on software-defined networking (SDN) and the design of its control plane is shown in Fig. 1(b). Hence, in the DCN, the servers are managed by the IT controller (IT-C) based on OpenStack for VM deployment and migration, while the switches (*i.e.*, both packet and optical ones) are controlled by the network controller (NET-C) based on ONOS to forward the traffic of computing tasks. The IT-C and NET-C are managed by a knowledge-defined orchestrator (KD-O), which makes wise decisions for network/IT resource orchestration. Then, based on the decisions, the system may invoke VM migration and/or network reconfiguration to provision the tasks cost-effectively. For instance, as suggested, the optical inter-rack network can be reconfigured in advance to prepare for future elephant flows, and in the meantime, VM migration can be triggered to organize the source and destination VMs of the tasks on proper racks and make the best use of the optical network. The H-O/E

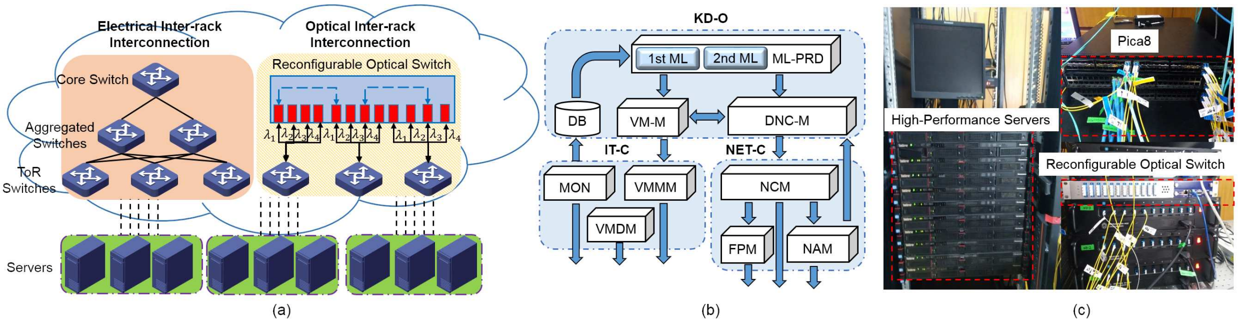

Electrical Inter-rack Interconnection
Optical Inter-rack Interconnection
Reconfigurable Optical Switch
Core Switch
Aggregated Switches
ToR Switches
Servers
λ1 λ2 λ3 λ4 λ1 λ2 λ3 λ4 λ1 λ2 λ3 λ4
(a)
KD-O
1st ML
2nd ML
ML-PRD
DB
VM-M
DNC-M
IT-C
NET-C
MON
VMMM
VMDM
NCM
FPM
NAM
(b)
High-Performance Servers
Pica8
Reconfigurable Optical Switch
(c)

Fig. 1. (a) Architecture of the H-O/E DCN, (b) Functional design of the control plane, and (c) Experimental testbed.

DCN system is implemented in a real network testbed as shown in Fig. 1(c), and we conduct experiments with live traffic on to demonstrate the knowledge-defined network/IT resource orchestration based on predictive analytics.

2. System Design and Operation Principle

The detailed design of the control plane in the H-O/E DCN system is shown in Fig. 1(b). In the IT-C, we implement two modules to realize dynamic VM deployment (VMDM) and migration (VMMM) in the servers, respectively, while an IT resource and traffic monitor (MON) is also implemented there to collect network statistics periodically and forward the results to the KD-O. We develop a network configuration module (NCM), a flow provisioning module (FPM), and a network abstraction module (NAM), and embed them in the NET-C. Specifically, the FPM is designed to groom the VMs' traffic into mice and elephant flows and route them in the electrical and optical inter-rack networks, respectively, and the NCM is controlled by the KD-O to configure the optical inter-rack network and update the topology in the NAM accordingly. The KD-O consists of an ML-based prediction and decision making module (ML-PRD), an IT resource and traffic database (DB), a VM management module (VM-M), and a DCN management module (DCN-M). Based on the historical traffic statistics in the DB, the ML-PRD first predicts future traffic and then determines a proper configuration of the optical inter-rack network to carry it. The decisions from the ML-PRD are implemented by the VM-M and DCN-M, which talk with the IT-C and NET-C, respectively, for orchestrating the network/IT resources.

Fig. 2 explains the principle of ML-PRD, which includes two ML modules and makes them work collaboratively to forecast future traffic precisely and then drive strategic decision making, *i.e.*, achieving predictive analytics.

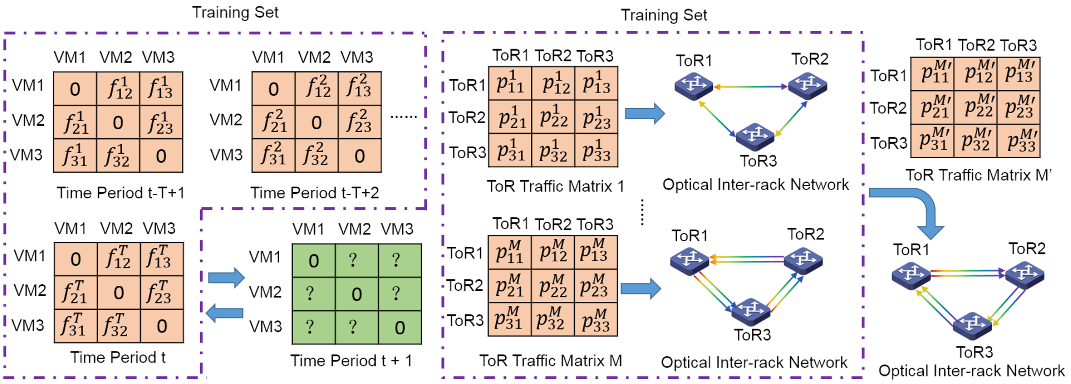
Predicting Future Traffic: As explained in Fig. 2(a), the first ML module uses the historical traffic matrixes among VMs in time periods $\{t-T+1,\cdots,t\}$ as the training set to predict the traffic matrix in time period $t+1$. This is repeated in every time period. Note that, as the training samples are unsupervised, we apply an enhanced ML method to insert newly-collected traffic matrixes in the training set consistently, for reducing the prediction loss.

Decision Making on Optical Network Configuration: With the predicted traffic matrix among VMs, the ML-PRD obtains the future traffic matrix among the top-of-rack (ToR) switches. Then, as indicated in Fig. 2(b), the second ML module determines the configuration of the optical inter-rack network with a supervised ML method. Specifically, we first set the module to its training phase, randomly generate enough combinations of ToR traffic matrix and optical network configuration, implement the combinations, and collect average data-transfer latency of the flows for each combination. Here, the average latency is used as the merit of an optical network configuration. We assume that all the flows are based on TCP, and the average latency is the average time that is spent by a task to finish its data-transfer between a VM pair. We use the combinations of ToR traffic matrix and optical network configuration and their average latencies to train the ML model. When the training has been done, the module is set to its operational phase. Hence, it can determine a proper optical network configuration based on the predicted traffic from the first ML module.

Finally, with the decisions from the second ML module, the ML-PRD chooses to either keep the H-O/E DCN system unchanged or reconfigure it accordingly, for realizing cost-effective network/IT resource orchestration.

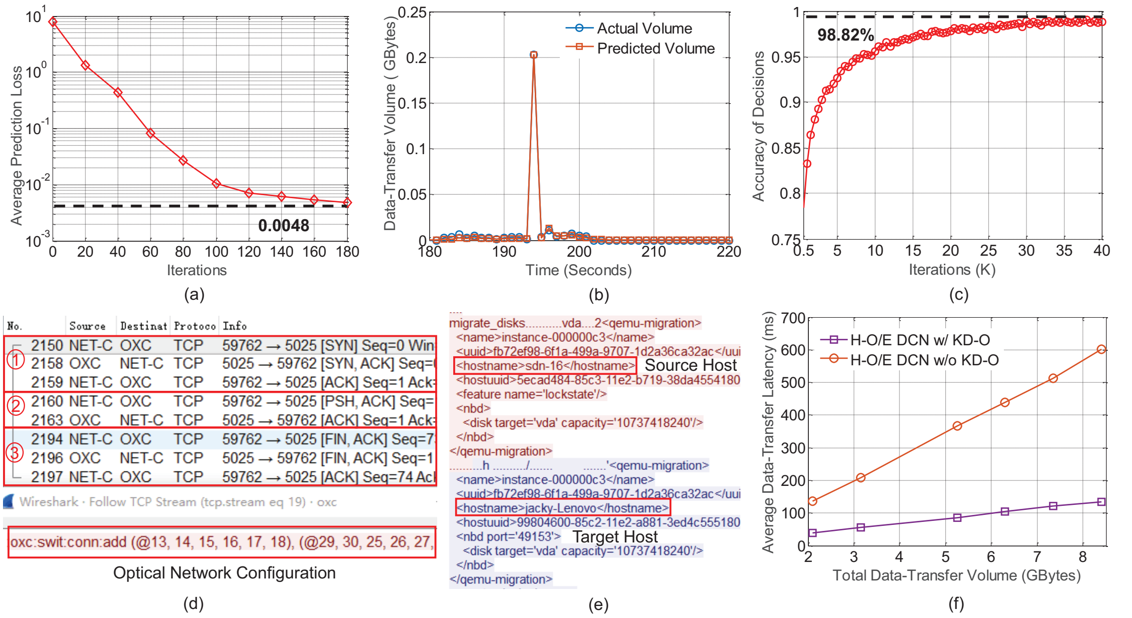
3. Experimental Demonstrations

Our experimental demonstrations use the network testbed shown in Fig. 1(a). The control plane of the H-O/E DCN, *i.e.*, the IT-C, NET-C and KD-O, is implemented in commodity servers. We leverage open-source softwares and modify them to fit into our requirements. Specifically, the IT-C and NET-C are based on OpenStack and ONOS, respectively, while the ML-based modules in the ML-PRD are implemented based on TensorFlow. The data plane of the H-O/E



(a) Predicting Future Traffic (b) Decision Making on Optical Network Configuration

Fig. 2. Operation principle of the ML-based prediction and decision making module (ML-PRD).



```
No.   Source Destinat Protoco Info
2150 NET-C  OXC      TCP     59762 → 5025 [SYN] Seq=0 Win
2158 OXC    NET-C    TCP     5025 → 59762 [SYN, ACK] Seq=
2159 NET-C  OXC      TCP     59762 → 5025 [ACK] Seq=1 Ack
2160 NET-C  OXC      TCP     59762 → 5025 [PSH, ACK] Seq=
2163 OXC    NET-C    TCP     5025 → 59762 [ACK] Seq=1 Ack
2194 NET-C  OXC      TCP     59762 → 5025 [FIN, ACK] Seq=7
2196 OXC    NET-C    TCP     5025 → 59762 [FIN, ACK] Seq=1
2197 NET-C  OXC      TCP     59762 → 5025 [ACK] Seq=74 Ac
```

```
oxc:swit:conn:add (@13, 14, 15, 16, 17, 18), (@29, 30, 25, 26, 27,
```

```
....
migrate_disks...........vda....2<qemu-migration>
  <name>instance-000000c3</name>
  <uuid>fb72ef98-6f1a-499a-9707-1d2a36ca32ac</uui
  <hostname>sdn-16</hostname>
  <hostuuid>5ecad484-85c3-11e2-b719-38da4554180
  <feature name='lockstate'/>
  <nbd>
    <disk target='vda' capacity='10737418240'/>
  </nbd>
</qemu-migration>
..........h ........../.......        .......'<qemu-migration>
  <name>instance-000000c3</name>
  <uuid>fb72ef98-6f1a-499a-9707-1d2a36ca32ac</uui
  <hostname>jacky-Lenovo</hostname>
  <hostuuid>99804600-85c2-11e2-a881-3ed4c555180
  <nbd port='49153'>
    <disk target='vda' capacity='10737418240'/>
  </nbd>
</qemu-migration>
```

Fig. 3. (a) Average prediction loss of the first ML module, (b) Example on predicted and actual data-transfer volumes between a VM pair, (c) Accuracy of decision making by the second ML module for optical network configuration, (d) and (e) Wireshark captures for optical network configuration and VM migration, and (f) Average data-transfer latency.

DCN system consists of four servers forming three racks to carry VMs, three ToR switches, six aggregated/core switches to build the electrical inter-rack network, and a reconfigurable optical switch (*i.e.*, Polatis 24×24 OXC) to realize the optical inter-rack network. Here, the ToR switches are commercial OpenFlow switches (*i.e.*, Pica-8). Each ToR switch is equipped with several 1GbE ports to connect with both the server(s) in its rack and the aggregate switches in the electrical inter-rack network, and it also has two optical 10GbE ports to realize inter-rack connections through the reconfigurable optical switch. In the experiments, we get the traffic matrixes with the DCT2GEN tool [5], and the obtained traffic matrixes are implemented by running iPerf on the VMs. Since we can hardly saturate the optical 10GbE ports on the ToR switches with the traffic generated by the four servers, we emulate the high traffic-load scenarios by limiting the ports' data-rates. Specifically, on each ToR switch, the 1GbE ports that connect to aggregated switches have a peak throughput of 300 Mbps, while the data-rate of each optical 10GbE port is limited below 700 Mbps.

Fig. 3 shows the experimental results. The average prediction loss in the first ML module, *i.e.*, the mean square error of the predicted data-transfer volumes among VMs to the actual ones, is presented in Fig. 3(a). We observe that our enhance ML method reduces the loss to 0.0048 (*i.e.*, a deviation ≤ 1 MByte) quickly within 180 iterations. Fig. 3(b) illustrates an example on the predicted and actual data-transfer volumes between a VM pair. The training performance of the second ML module is plotted in Fig. 3(c), which indicates that after 4×10^4 iterations (*i.e.*, only taking a few seconds), the module can provide the best decisions with an accuracy of 98.82%, *i.e.*, it can find the optimal optical network configuration automatically with a high accuracy. To verify the feasibilities of network/IT resource orchestration in the H-O/E DCN, Figs. 3(d) and 3(e) shows the wireshark captures for optical network reconfiguration and VM migration, respectively. As indicated in Fig. 3(d), the NET-C sets up a TCP connection to operate the reconfigurable optical switch (*i.e.*, OXC). The message in Fig. 3(e) suggests that a VM at host "sdn-16" has migrated to host "jacky-Lenovo" successfully. Fig. 3(f) compares the average data-transfer latency of flows in the H-O/E DCN for with and without the KD-O, and the results verify that the latency is reduced effectively with our proposed KD-O.

4. Conclusion

We follow the principle of predictive analytics to design and experimentally demonstrate an H-O/E DCN system that achieves knowledge-defined network/IT resource orchestration to reduce the average data-transfer latencies effectively.

References

[1] G. Wang *et al.*, in *Proc. of SIGCOMM 2010*, pp. 327-338, Sept. 2010.
[2] C. Kachris *et al.*, *IEEE Commu. Surveys Tut.*, vol. 14, no. 4, pp. 1021-1036, Jan. 2012.
[3] N. Hamedazimi *et al.*, in *Proc. of SIGCOMM 2014*, pp. 319-330, Aug. 2014.
[4] T. Benson *et al.*, in *Proc. of SIGCOMM 2010*, pp. 267-280, Nov. 2010.
[5] P. Wette *et al.*, *arXiv:1409.2246*, Sept. 2014.