NetO: Alibaba's WAN Orchestrator

Xin Wu Chao Huang Ming Tang Yihong Sang Wei Zhou Tao Wang Yuan He Dennis Cai Haiyong Wang Ming Zhang
{x.wu, jingtan.hc, ming.tang, yihong.syh, w.zhou, junchen.wt, jinghang.hy, d.cai, haiyong.wang, m.zhang}@alibaba-inc.com

1. ABSTRACT

Alibaba has one of the largest enterprise WANs in the Internet. It connects various datacenters and ISPs across the globe. We present the design of Alibaba's Intelligent WAN Orchestration System (*i.e.*, NetO). The key challenge for NetO is to satisfy the bandwidth and latency requirements from applications of different priorities under dynamic network conditions and traffic demands without overprovisioning the expensive long-haul capacity. NetO achieves high utilization and low latency by both traffic engineering across datacenters and peering engineering across ISPs.

2. INTRODUCTION

Alibaba has one of the largest enterprise WANs in the Internet. It serves a wide range of e-commerce, cloud, video, and enterprise applications and inter-connects various datacenters and ISPs. Unlike the previous work where inter-datacenter traffic engineering and ISP peering are managed by separate systems [1–3], we design NetO to manage both. On the one hand, the WAN capacities are carefully allocated according to the traffic priorities under dynamic network conditions and traffic demands to achieve high utilization. On the other hand, the ISP peering policies are frequently updated to achieve low latency between datacenters and end users.

NetO's design goal has two-fold: First, it supports flexible traffic priority classification and allocates bandwidth in strict precedence of these priorities, while preferring paths with lower latencies for higher priorities. Second, within the same priority, it allocates bandwidth to achieve max-min fairness. From a bird's eye view, NetO coordinates traffic demands, network conditions and operators' intentions, computes the bandwidth allocation and updates the routers' forwarding states. Despite the conceptual simplicity, we must address two challenges to fulfill this design.

First, we need a scalable algorithm to compute a congestion-free, multi-step transition plan. Each step involves multiple router updates, which may change either the traffic distribution or the ISP peering policies. Irrespective of the asynchronous nature of the updates, each step must subject to constraints on capacity, traffic priority and fairness. To resolve this challenge, we formalize the problem using an optimization model and develop a heuristics to approach an approximation. We then translate the approximation to configuration changes, BGP updates and tunnels with explicit paths, which are further applied to routers.

Second, for incremental deployment and fault tolerance purpose, we would like that NetO starts from taking a small portion of the WAN traffic and gradually increases the portion. Even for this small portion, we would like that the WAN can degenerate to the legacy forwarding when NetO fails. We resolve this challenge by designing NetO as an "overlay" on top of the existing forwarding. We rely on the robustness of a link state IGP protocol to manage the topology adjacency. On top of the IGP, BGP is used to propagate prefix reachability. These two layers form the basis of NetO. NetO optimizes paths for a portion of the traffic and overwrites the default paths provided by the underlying IGP and BGP. We choose segment routing to implement NetO not only because it naturally fits in this layered architecture, but also because path updating in segment routing touches only the head of the tunnel and thus simplifies the implementation of control plane consistency [4].

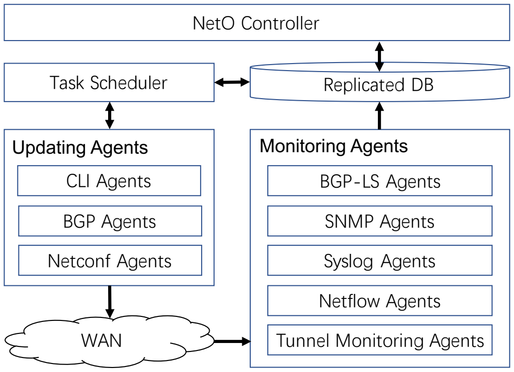

Figure 1: NetO Architecture

3. DESIGN

Figure 1 shows the architecture of NetO. A replicated database serves as a persistent store for all the components, such that these components remain stateless and can recover from crashes. A centralized controller uses the topology, the traffic demand and operators' intentions to periodically compute the ISP peering policies and the WAN bandwidth allocation. The task scheduler is responsible to schedule the controller decisions to various updating agents based on the task types, priorities, security constrains and workloads. There are two types of agents that directly talk to WAN routers: the updating agents that can change router forwarding states and the monitoring agents that collect information from the routers. Each type of agent has its own way of deployment to achieve consistency and high availability. For example, all updating agents are deployed as active; BGP-LS agents which collect topology information are deployed as active-standby such that only the master agent is updating the topology; Netflow agents are actually deployed in hierarchies, such that traffic demand is properly aggregated before writing to the DB.

4. CONCLUSION

This work presents the design of Alibaba's WAN orchestrator NetO. It achieves high utilization and low latency by both traffic engineering across datacenters and peering engineering across ISPs. The core of NetO is an optimization model that computes a congestion-free transition plan. To minimize the deployment disruption, we have implemented and incrementally deployed NetO using segment routing on top of the existing IGP and BGP.

5. REFERENCES

[1] R. Hartert, S. Vissicchio, P. Schaus, O. Bonaventure, C. Filsfils, T. Telkamp, and P. Francois. A declarative and expressive approach to control forwarding paths in carrier-grade networks. *SIGCOMM Comput. Commun. Rev.*, 45(4), Aug. 2015.

[2] C.-Y. Hong, S. Kandula, R. Mahajan, M. Zhang, V. Gill, M. Nanduri, and R. Wattenhofer. Achieving high utilization with software-driven wan. *SIGCOMM Comput. Commun. Rev.*, 43(4), Aug. 2013.

[3] S. Jain, A. Kumar, S. Mandal, J. Ong, L. Poutievski, A. Singh, S. Venkata, J. Wanderer, J. Zhou, M. Zhu, J. Zolla, U. Hölzle, S. Stuart, and A. Vahdat. B4: Experience with a globally-deployed software defined wan. *SIGCOMM Comput. Commun. Rev.*, 43(4), Aug. 2013.

[4] M. Reitblatt, N. Foster, J. Rexford, C. Schlesinger, and D. Walker. Abstractions for network update. In *Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, SIGCOMM '12. ACM, 2012.