(Invited) How Traffic Analytics Shapes Traffic Engineering, Topology Engineering, and Capacity Planning of Jupiter

Anny Xijia Zheng, Jianan Zhang, Rui Wang, Leon Poutievski
Google LLC, 1600 Amphitheatre Pkwy, Mountain View, CA 94043 USA
{*axzheng, jiananzhang, ruiw, leonp*}*@google.com*

Abstract: Three prominent traffic features including peak alignment, stable ranking, and gravity model, have guided the design of current Google Jupiter fabrics in traffic engineering, topology engineering, and capacity planning.

1. Introduction

We have witnessed the fast growth of datacenter network traffic from various network services and applications in the recent two decades. Google's aggregated regular server traffic has multiplied 235 times from 2011 to 2021 [1]. Furthermore, traffic from artificial intelligence and machine learning applications grew even faster in the last decade. The number of unique project directories related to machine learning increased fivefold to 7,500 between 2015 and 2017 [2]. All of these require more efficient networking infrastructure to support more advanced computing power.

Google's Jupiter datacenter network fabrics are highly available systems designed to support data communications within the large-scale distributed computing systems. Each modern Jupiter datacenter network fabric consists of machine aggregation blocks which are directly and dynamically connected by optical circuit switches to support communications between pairs among tens of thousands of servers. To further reflect the application communication patterns and speed heterogeneity [3], traffic engineering adapts the routing scheme timely and topology engineering reconfigures mesh connectivity dynamically. Before this, Jupiter fabrics used Clos topology for a decade [4]. The legacy three-tier Clos topology involves spine switches interconnecting the aggregation blocks to provide uniform bandwidth across all the servers in the same fabric. The motivations behind this evolution are the observations on the performance bottleneck of spines, traffic pattern analysis, and power consumption reduction [3]. In this paper, we focus on the traffic pattern analysis and illustrate three prominent traffic features, including peak alignment, stable ranking, and gravity model. We discuss the impacts of these three features on shaping the designs of traffic engineering, topology engineering, and capacity planning for today's Jupiter fabrics.

2. Traffic Measurement and Representation

Google's network monitoring and telemetry capabilities [5] on live production traffic are the foundations behind traffic analytics. The traffic measurement pipeline gathers fine-grained datacenter network traffic information, including demands, bandwidth usage, end-to-end performance, etc. Traffic can be collected vertically in server level or block level, and horizontally in service and application level. For example, we collect per-server flow measurements through packet sampling or flow counter every 30 seconds for traffic-aware routing in traffic engineering [3]. The traffic storage and query pipelines support detailed filtering and representation. Traffic information can be acquired and organized with various criteria, such as categories, spatial order, temporal order, etc.

Traffic is represented in the format of a time series of *traffic matrices*. Each traffic matrix consists of traffic from source aggregation block to destination aggregation block in the same datacenter fabric. A *commodity* is defined as the traffic flow from a source block to a destination block. A traffic matrix contains all commodities' traffic. In the following sections, we present three prominent features of the traffic matrix time series and explain how they shape the system designs.

3. Traffic Characteristics

Datacenter network traffic has variations in both temporal and spatial dimensions. In the temporal dimension, traffic exhibits seasonal patterns and is volatile under the 30-second sampling rate. In the spatial dimension, some commodities have larger traffic volumes than others, and traffic of different commodities follows a skewed distribution. To support different users and applications, different datacenters generate traffic with different magnitudes of volatility and skewness. There are three prominent features that are universal over existing datacenters and

make traffic and topology engineering in direct-connect topology viable. First, most commodities have correlated traffic volumes and the traffic peaks occur almost at the same time. Second, the rankings of commodity traffic volumes are stable. Third, inter-aggregation-block commodity traffic can be described by the aggregation block's total traffic using the gravity model. We first illustrate these features in this section, and then discuss how these features guide the system design in the next section.

3.1. Peak alignment

Traffic from different aggregation blocks are correlated as a result of the fact that datacenters support distributed computing over a large number of servers under multiple aggregation blocks. The peaks of traffic from different aggregation blocks align almost at the same time. Figure 1 illustrates the egress traffic from each aggregation block in a fabric during a week. Moreover, the peaks of commodity traffic also align almost at the same time.

Let $T_{ij}(t)$ denote the traffic volume from aggregation block i to j at time t. Let $T^*_{ij} = \max_{t\in T} T_{ij}(t)$ denote the maximum commodity-(i,j) traffic over time interval T. Let $T(t) = \sum_{i,j\in V} T_{ij}(t)$ denote the total traffic between all aggregation blocks in a datacenter, where V denotes the set of aggregation blocks. Let $T^* = \max_{t\in T} T(t)$ denote the maximum total traffic over T. If all commodity peaks are aligned during T, then $R = T^*/\sum_{i,j\in V} T^*_{ij} = 1$. Smaller R implies that peaks are asynchronous. The range of R is within $(0.7, 0.9)$ for ten typical datacenters for $T = 24$ hours. The range of R decreases to $(0.6, 0.85)$ for $T = 30$ days. Therefore, commodity peaks are well aligned even during a long time interval. Taking the maximum over a time interval for each commodity well characterizes the worst-case traffic matrix, under which link utilizations are high and the network is most likely congested.

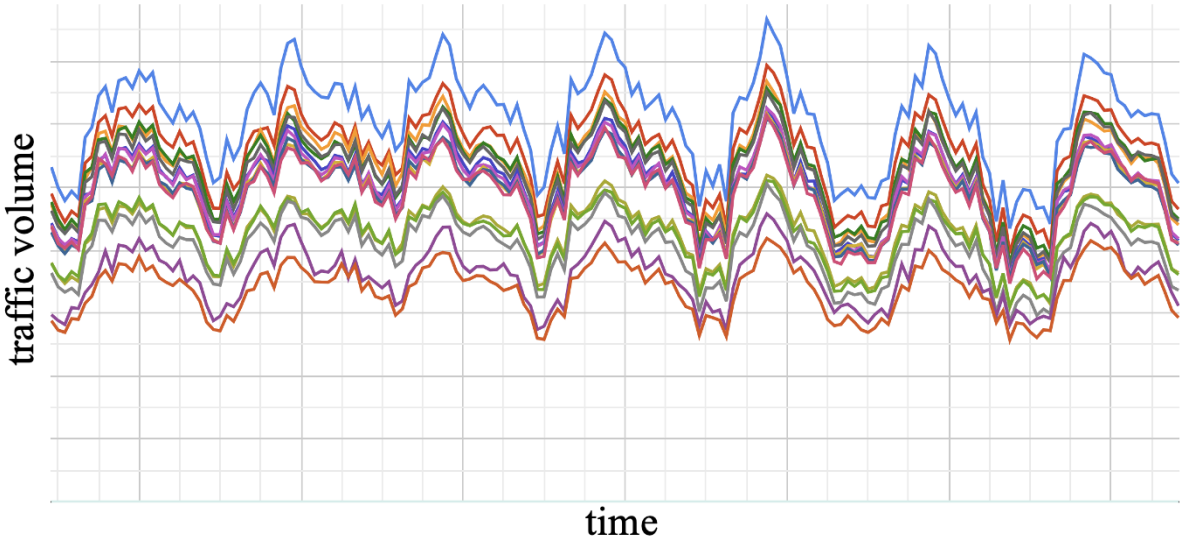

Fig. 1: Traffic from each aggregation block for a fabric during a week.

3.2. Stable ranking

While traffic from different aggregation blocks fluctuates, the rankings of block traffic volumes remain stable as shown in Fig. 1. Exception occurs when there are more machines turned up under a block and its traffic increases. Since traffic volume rankings are stable, congestion usually occurs at a fixed set of aggregation blocks. Therefore, it is plausible to upgrade the congested blocks based on the past traffic. Moreover, topology engineering can use past traffic to compute a topology to support similar future traffic.

3.3. Gravity model

Uniform random machine-to-machine communications generate traffic that follows the gravity model between each aggregation block pair. Let E_i denote the egress traffic from aggregation block i and I_j denote the ingress traffic to aggregation block j. Let $T = \sum_{i\in V} E_i = \sum_{j\in V} I_j$ denote the total traffic. Under the gravity model, traffic from i to j is $E_i I_j / T$, which states the demand from node i to node j is proportional to the product of the egress demand at i and the ingress demand at j. The gravity model provides the traffic predictability to reduce the topology needs for Clos topology.

4. System Design Based on Traffic Characterizations

4.1. Traffic Engineering

Peak alignment contributes to the current traffic engineering design. Traffic is routed over multiple paths using traffic engineering. To compute the traffic split ratios over multiple paths, we formulate a multi-commodity flow problem to minimize the maximum link utilization under a predicted traffic matrix T^*. Since commodity traffic peaks are aligned, T^* is predicted by taking the maximum traffic for each commodity over a time interval that is similar to the routing reconfiguration interval. Under any fixed routing, the link utilizations under traffic matrix $T(t)$ will not exceed the link utilizations under $T^* \geq T(t)$. Therefore, it is sufficient to apply the same routing for T^* to support $T(t)$. We observe that hourly reconfiguration is sufficient to support the dynamic traffic.

4.2. Topology Engineering

Direct-connect topology is able to support the same set of traffic matrices as Clos topology if inter-block traffic follows the gravity model [3]. Furthermore, traffic predictability provided by the three traffic features enable us to adopt non-uniform direct-connect topology to efficiently support datacenter traffic, by constructing a larger capacity between two blocks with larger traffic between them. Since block traffic volumes have stable rankings over weeks and even months, a static topology that is designed for the past traffic works well for the future. On occasions when traffic shifts due to newly landed machines or changing workloads, the topology can be reconfigured to support new traffic patterns.

4.3. Capacity Planning

The relatively predictable and stable traffic peaks during a long time interval simplify datacenter network capacity planning, which includes upgrading block radices and adding new blocks. Capacity planning usually takes several months from proposal to deployment. In the proposal stage, different proposals to increase the network capacity and reduce congestion are evaluated using past traffic matrices. Since traffic rankings are stable, reducing the utilizations of congested blocks by upgrading radices to increase capacity will avoid block congestion for a long time interval. By reducing link utilizations to a threshold that allows sufficient headroom, the proposed capacity augmentation will be able to support future traffic growth over the next few months.

5. Conclusion and Future Work

Google's current Jupiter datacenter network fabrics have been in production for more than five years. The transformation from Clos topology to direct-connect topology powered by traffic analytics enhances the flexibility, scalability, stability, and availability of such large production networks. There are three main observations on the inter-block traffic: peak alignment, stable ranking, and gravity model. They shape the design of traffic engineering, topology engineering, and capacity planning of datacenter networks.

In the future, we plan to further develop traffic analytics as a powerful tool to approach the network performance upper bound. The diverse traffic patterns of different applications provide us with new challenges and opportunities. In particular, we hope to explore more on both traffic engineering and topology engineering to support the new traffic patterns brought by the fast growing artificial intelligence and machine learning applications.

References

1. C. Lam, X. Zhou, and H. Liu, "The path towards 3.2T pluggable datacenter transceivers," (2022). Optical Fiber Communication Conference and Exhibition (OFC) Market Watch Panel on Building the Next Generation 3.2T Transceiver.
2. L. A. Barroso, U. Hölzle, and P. Ranganathan, "The datacenter as a computer: Designing warehouse-scale machines, third edition," Synthesis Lectures on Computer Architecture **13**, i–189 (2018).
3. L. Poutievski, O. Mashayekhi, J. Ong, A. Singh, M. Tariq, R. Wang, J. Zhang, V. Beauregard, P. Conner, S. Gribble, R. Kapoor, S. Kratzer, N. Li, H. Liu, K. Nagaraj, J. Ornstein, S. Sawhney, R. Urata, L. Vicisano, K. Yasumura, S. Zhang, J. Zhou, and A. Vahdat, "Jupiter evolving: Transforming Google's datacenter network via optical circuit switches and software-defined networking," in *Proceedings of the ACM SIGCOMM 2022 Conference,* (Association for Computing Machinery, New York, NY, USA, 2022), SIGCOMM '22, p. 66–85.
4. A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai, B. Felderman, P. Germano, A. Kanagala, J. Provost, J. Simmons, E. Tanda, J. Wanderer, U. Hölzle, S. Stuart, and A. Vahdat, "Jupiter rising: A decade of clos topologies and centralized control in Google's datacenter network," in *Proceedings of the ACM SIGCOMM 2015 Conference,* (2015), SIGCOMM '15.
5. "Deploy network monitoring and telemetry capabilities in google cloud," https://cloud.google.com/architecture/deploy-network-telemetry-blueprint. Accessed: 2023-01-20.