International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(20s) 2026 pp. 732-740
https://doi.org/10.70917/ijcisim-2026-5216
Article

Cisco NDFC-Managed Spine-and-Leaf Architecture: An Analytical Model for Deterministic, Automated VXLAN EVPN Data-Center Fabrics

Chetan Padshala

Campbells University, Kentucky
ORCID: 0009-0002-5966-4103

Abstract: Data-center workloads driven by artificial-intelligence training, cloud-native services, and distributed microservices have outgrown the three-tier network designs that once served enterprise applications. Hierarchical aggregation with the Spanning Tree Protocol wastes capacity on blocked links and introduces variable path lengths that modern east-west traffic cannot tolerate. Spine-and-leaf fabrics built on the Clos principle answer these challenges by offering equal-cost multipath forwarding, predictable hop counts, and horizontal scale-out, and VXLAN with a BGP EVPN control plane extends multi-tenant Layer-2 semantics across a routed underlay. Operating such a fabric by hand does not scale, so controller-based automation has become central to the design rather than an accessory. This paper analyzes the Cisco Nexus Dashboard Fabric Controller as the automation layer of a spine-and-leaf VXLAN EVPN fabric and develops a compact analytical model that links the number of spine switches to equal-cost path diversity, oversubscription ratio, bisection bandwidth, and fabric availability. The model is evaluated across representative fabric configurations, showing how spine count sets both the performance envelope and the resilience of the fabric and how a non-blocking ratio is reached at a determinable spine count. A mapping of controller functions to fabric lifecycle stages clarifies where automation reduces configuration drift and operational risk. The analysis indicates that deterministic performance at scale is an emergent property of three decisions taken together, namely a Clos topology, an EVPN control plane, and controller-enforced automation, and that treating any one in isolation understates the design. Migration and operational-expertise trade-offs are discussed to frame practical adoption.
Keywords: Spine-and-Leaf, Clos Fabric, VXLAN EVPN, Data-Center Networking, Network Automation, ECMP, Deterministic Latency

1. Introduction

Modern data centers carry a traffic mix that bears little resemblance to the client-server patterns for which legacy networks were built. Distributed training of large models, cloud-native application meshes, and microservice architectures generate dense east-west communication in which any server may exchange high-bandwidth flows with any other. The three-tier design of access, aggregation, and core, held loop-free by Spanning Tree Protocol, was never intended for this regime; it blocks redundant links to prevent loops, forces traffic through aggregation choke points, and produces path lengths that vary with the position of the endpoints [4].

The remedy has deep theoretical roots. Non-blocking multistage switching was formalized in the 1950s, when it was shown that a multistage fabric can connect inputs to outputs without internal contention if its stages are dimensioned correctly [1]. Data-center architects rediscovered the principle in the form of fat-tree and Clos topologies built from commodity switches [2], and in flexible Layer-2 designs that preserved uniform capacity across a large server population [3]. The two-tier spine-and-leaf fabric is the practical realization of these ideas: every leaf connects to every spine so that any leaf-to-leaf path crosses exactly one spine and all such paths are equal in cost [6].

Topology alone does not deliver a usable fabric. A routed underlay must carry encapsulated tenant traffic, and a control plane must distribute endpoint reachability without the flooding that plagued earlier Layer-2 networks. Virtual Extensible LAN provides the encapsulation that stretches Layer-2 segments across a Layer-3 fabric [7], and a BGP Ethernet VPN control plane advertises MAC and IP endpoints so that forwarding is learned rather than flooded [8]. The combination gives multi-tenant segmentation and scalable reachability, but it also multiplies the configuration surface, since every leaf and spine must be provisioned consistently for the fabric to behave deterministically.

Consistency at that scale is a problem of automation rather than of protocol. Software-defined control separates policy from device configuration, making centralized management feasible [9], and controller platforms now translate an operator's intent into synchronized device configuration across a fabric [10], [11]. The Cisco Nexus Dashboard Fabric Controller occupies this role for Nexus-based fabrics, provisioning underlay and overlay, enforcing templates, and detecting drift from a single control point. The question this paper examines is how these three decisions, a Clos topology, an EVPN control plane, and controller automation, combine to produce a fabric that is simultaneously deterministic, scalable, and operable.

The contributions are fourfold. First, an analytical model is developed that relates spine count to equal-cost path diversity, oversubscription ratio, and bisection bandwidth so that a fabric can be dimensioned to a performance target rather than by rule of thumb. Second, a parallel-redundancy availability model quantifies how spine multiplicity governs fabric resilience. Third, the controller's automation functions are mapped to the stages of the fabric lifecycle, locating where automation reduces drift and operational risk. Fourth, migration and operational-expertise trade-offs are discussed to frame adoption.

2. Related Work

Research on data-center networking has matured along three lines that this paper draws together. The first concerns topology. Surveys of data center network infrastructure classify designs into switch-centric and server-centric families and identify the Clos-derived spine-and-leaf fabric as the prevailing switch-centric choice for its uniform capacity and scale-out behavior [4]. The performance of the leaf-spine data path has been characterized directly, confirming that a well-dimensioned two-tier fabric approaches non-blocking behavior under realistic traffic [6]. The trajectory of the field has since turned from low average latency toward bounded, deterministic latency as a first-class objective [5].

The second line concerns the control and overlay planes. The VXLAN framework defined the encapsulation used to carry Layer-2 segments over a routed fabric [7], and the BGP EVPN specification supplied a control plane that advertises endpoint reachability and supports multi-tenant segmentation [8]. Practical fabric builds combine the two, and recent work has automated their provisioning, including decentralized and dynamic zero-touch provisioning of leaf-spine EVPN fabrics [11]. Software-defined networking underpins this shift by separating control from forwarding [9], and intent-based networking frames the operator interface as declarative policy that a controller renders into device state [10].

The third line concerns performance management inside the fabric. Load balancing across equal-cost paths is a persistent concern, since naive hashing can concentrate flows and erode the benefit of path diversity; surveys map the design space from static ECMP to congestion-aware schemes [12], [14], and distributed congestion-aware balancing was shown to improve utilization in Clos fabrics [13]. Programmable data planes have extended what the fabric can measure and enforce, from P4-based switches [15] to AI-driven forwarding [16] and fine-grained in-band telemetry [17]. For the lossless behavior that distributed training demands, buffer management and congestion control for lossless Ethernet have been re-examined [18], [19], and large operators have documented purpose-built fabrics for model training at scale [20], [21]. Reliability modeling completes the picture by quantifying availability and redundancy in data-center topologies [22]. The present work distinguishes itself by treating topology, control plane, and automation as one analytical object rather than as separate concerns.

3. Materials and Methods

3.1 Spine-and-leaf fabric principles

A spine-and-leaf fabric is a two-tier folded Clos network. Leaf switches attach servers and provide Layer-2 and Layer-3 edge services; spine switches provide high-bandwidth Layer-3 transit between leaves, and every leaf connects to every spine. Because a leaf-to-leaf flow always traverses exactly one spine, hop count is uniform, and the equal-cost paths between any pair of leaves number exactly as many as there are spines. Redundant links carry traffic simultaneously rather than being blocked, which removes the central inefficiency of the Spanning Tree Protocol and allows the fabric to scale by adding switches rather than replacing larger ones. The symbols used in the analytical model are listed in Table 1.

Table 1. Notation used in the fabric model.

Symbol	**Definition**
S	Number of spine switches
L	Number of leaf switches
R_up	Per-uplink line rate (leaf to spine)
D_down	Total server-facing (downlink) capacity per leaf
p_down	Server-facing ports per leaf
R_spine	Spine port radix (leaf-facing ports per spine)
a	Availability of a single spine path
OSR	Oversubscription ratio (downlink to uplink)

Table 2 contrasts the legacy three-tier design with the spine-and-leaf fabric on the dimensions that matter for east-west workloads.

Table 2. Three-tier hierarchy versus spine-and-leaf Clos fabric.

Property	**Three-tier (STP)**	**Spine-and-leaf (Clos)**
Redundant links	Blocked by STP	Active via ECMP
Path length	Variable (up to core)	Uniform (one spine hop)
Scaling	Vertical (bigger boxes)	Horizontal (add switches)
Default gateway	Centralized	Distributed anycast on leaves
Layer-2 extension	VLAN-bounded	VXLAN over routed underlay
Control plane	Flood-and-learn	BGP EVPN

3.2 Underlay and overlay

The underlay is a routed Layer-3 network whose sole task is to carry encapsulated overlay traffic between leaf and spine loopback addresses. A link-state protocol such as OSPF or IS-IS, or alternatively BGP, provides loop-free reachability and exposes the equal-cost paths that give the fabric its multipath behavior. The overlay encapsulates tenant Layer-2 frames in VXLAN and carries them across that underlay [7]. Endpoint reachability in the overlay is distributed by a BGP EVPN control plane, in which spines commonly act as route reflectors so that leaves need not peer in a full mesh [8]. Forwarding is therefore control-plane driven, with MAC and IP endpoints advertised as routes rather than discovered by flooding, and tenant isolation is expressed through virtual routing instances and VXLAN network identifiers.

3.3 Controller-based automation model

Manual configuration of a fabric with dozens of switches, each requiring consistent underlay, overlay, and policy state, is slow and error-prone, and configuration drift between devices is a common source of outages. A fabric controller addresses this challenge by holding the intended state centrally and rendering it into device configuration. Table 3 maps the controller functions considered here to the lifecycle stage each serves.

Table 3. Controller automation functions mapped to fabric lifecycle stages.

Lifecycle stage	Controller function	Effect
Day 0 (build)	Automated underlay provisioning (OSPF/IS-IS/BGP)	Consistent routed fabric
Day 0 (build)	Automated overlay provisioning (VXLAN EVPN)	Uniform overlay state
Day 1 (deploy)	Template- and intent-based configuration	Reduced human error
Day 1 (deploy)	Leaf/spine onboarding automation	Faster scale-out
Day 2 (operate)	Compliance and drift detection	Early fault avoidance
Day 2 (operate)	Policy-driven segmentation	Consistent multi-tenancy

The mapping shows automation acting across the whole lifecycle rather than only at build time, and it locates the operational value of the controller in the reduction of drift and in the enforcement of a single intended state. This is consistent with the direction of intent-based networking, in which declarative policy is continuously reconciled against the running fabric [10], and with demonstrated zero-touch provisioning for EVPN leaf-spine fabrics [11]. The proposed architecture is shown in Figure 1.

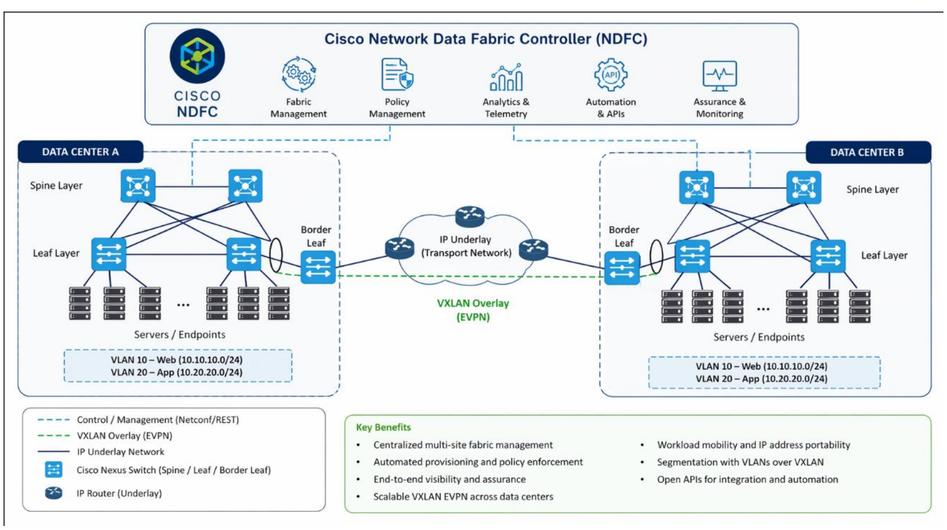

Fig 1. Controller-managed spine-and-leaf VXLAN EVPN fabric: routed underlay, EVPN overlay, distributed anycast gateways on leaves, and centralized lifecycle management.

3.4 Analytical fabric model

The performance envelope of a spine-and-leaf fabric can be expressed in closed form. The oversubscription ratio of a leaf, the balance between the capacity it offers to servers and the capacity it has toward the spines, is given by Equation (1), where a ratio at or below unity denotes a non-blocking leaf.

$$OSR = D\_down / (S * R\_up) \qquad (1)$$

The number of equal-cost paths between any two leaves equals the number of spines, as in Equation (2), which is the source of both the load-balancing capacity and the path redundancy of the fabric.

$$N\_paths = S \qquad (2)$$

Server capacity scales with the leaf count and the server-facing ports per leaf, as in Equation (3), while the leaf count is bounded by the spine port radix because each spine must reach every leaf, as in Equation (4).

$$N\_servers = L * p\_down \qquad (3)$$

$$L <= R\_spine \,, \quad uplinks\ per\ leaf = S \qquad (4)$$

The bisection bandwidth, a standard measure of the fabric's ability to sustain all-to-all traffic, is derived from the total leaf-to-spine capacity, as shown in Equation (5).

$$BW\_bisect = (L * S * R\_up) / 2 \qquad (5)$$

Fabric resilience is treated with the same parallel-redundancy reasoning used for redundant paths in general. Since the S spine paths between two leaves fail independently to a first approximation, the fabric path between them is available whenever at least one spine path survives, as in Equation (6), where a is the availability of a single spine path.

$$A\_fabric = 1 - (1 - a)^S \qquad (6)$$

Equations (1) through (6) make spine count the primary design lever: raising S lowers oversubscription, adds equal-cost paths, widens bisection bandwidth, and improves availability, at the cost of additional spine hardware and leaf uplink ports. The results section evaluates these relations for a representative leaf configuration.

4. Results

The model was evaluated for a representative leaf offering 1,200 Gbps of server-facing capacity, corresponding to forty-eight 25-Gbps downlinks, with 100-Gbps uplinks to the spines. Table 4 applies Equations (1) and (2) across spine counts from two to sixteen, reporting uplink capacity, oversubscription ratio, and equal-cost path count.

Table 4. Computed oversubscription and path diversity versus spine count (48x25G downlinks, 100G uplinks).

Spines S	Uplink capacity (Gbps)	Downlink capacity (Gbps)	Oversubscription (down:up)	ECMP paths
2	200	1,200	6.0:1	2
4	400	1,200	3.0:1	4
6	600	1,200	2.0:1	6
8	800	1,200	1.5:1	8
12	1,200	1,200	1.0:1 (non-blocking)	12
16	1,600	1,200	0.75:1	16

The table locates the non-blocking point precisely: with this leaf, twelve spines equalize downlink and uplink capacity, and beyond that, the fabric is undersubscribed. Path diversity rises in lockstep with spine count, so the same lever that removes oversubscription also multiplies the equal-cost paths available to load balancing. Figure 2 plots both quantities and shows the oversubscription curve crossing the non-blocking line as the path count grows linearly.

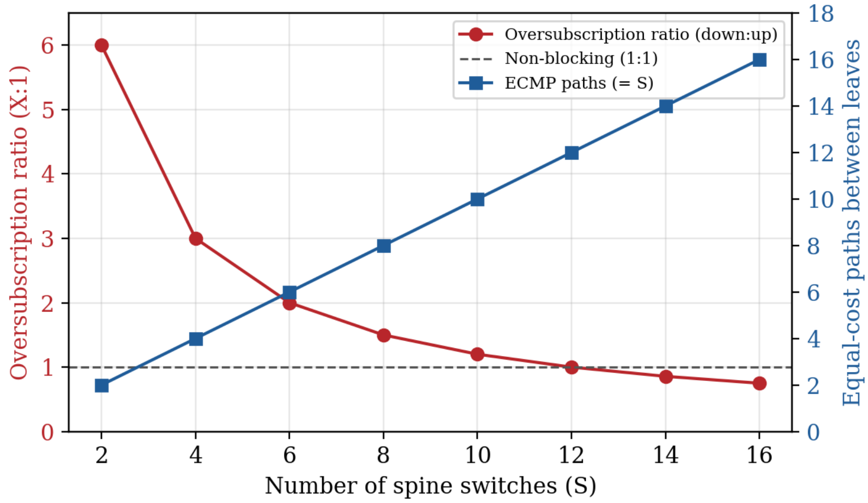

Fig 2. Oversubscription ratio and equal-cost path count as functions of spine count, computed from Equations (1) and (2). The non-blocking ratio is reached at twelve spines for the modeled leaf.

Resilience was evaluated using Equation (6) for three values of single-path availability. Table 5 reports the fabric path availability as spine count grows, and Figure 3 shows the same relationship. Even modest per-path availability yields high fabric availability once several spines are present, because the probability that all parallel paths fail simultaneously falls geometrically with spine count.

Table 5. Computed fabric path availability versus spine count, from Equation (6).

Spines S	a = 0.90	a = 0.95	a = 0.99
2	0.990000	0.997500	0.999900
4	0.999900	0.999994	0.99999999
6	0.999999	0.99999998	~1.0
8	~1.0	~1.0	~1.0

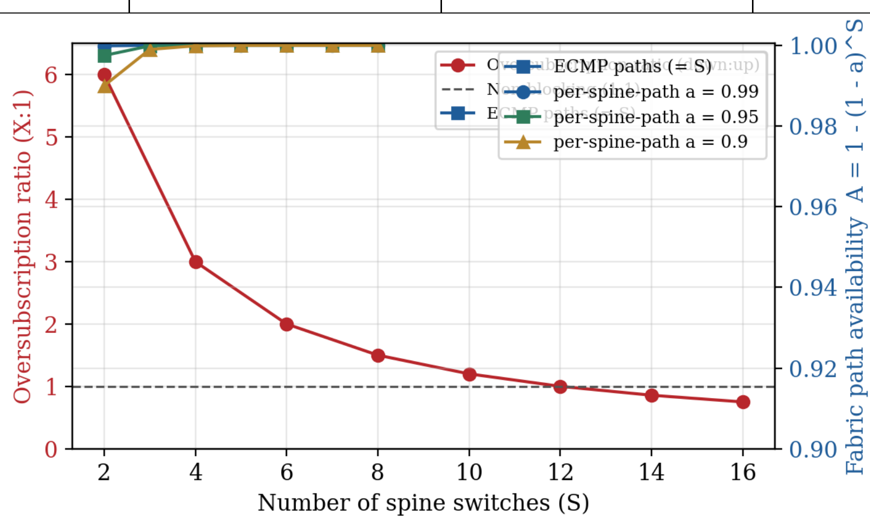

Fig 3. Fabric path availability as a function of spine count for three single-path availabilities, computed from the parallel-redundancy model of Equation (6).

Read together, the two results show that spine count is a single lever acting on performance and resilience at once. A fabric dimensioned for a non-blocking ratio already carries enough spines to place fabric availability well beyond the level any single switch could provide, which means that the performance target and the resilience target are not in tension but are met by the same design decision. Bisection bandwidth from Equation (5) scales with the product of leaf and spine counts, so capacity grows with the fabric rather than being fixed by a core tier.

5. Discussion

The analysis supports a specific reading of why controller-automated spine-and-leaf fabrics have displaced hierarchical designs. Determinism is not a property of any single element. The Clos topology supplies uniform path length and equal-cost diversity, the EVPN control plane supplies flood-free reachability and tenant isolation, and controller automation supplies the configuration consistency without which the first two degrade into drift and outage. Treating the topology in isolation, as a purely physical matter, understates the role of the control and automation planes in making the physical fabric behave predictably at scale.

The model also clarifies a design economy. Because spine count simultaneously governs oversubscription, path diversity, bisection bandwidth, and availability, a single dimensioning decision propagates across all four properties, and Table 4 and Table 5 let that decision be made against explicit targets. This choice matters most for the traffic that motivates the design. Distributed training and high-performance computing generate synchronized all-to-all traffic that punishes both oversubscription and path collisions, which is why operators have built dedicated fabrics for model training and why lossless-Ethernet behavior has received close attention [18], [19], [20], [21]. Congestion-aware load balancing across the equal-cost paths, rather than static hashing, extracts the diversity that Equation (2) provides [12], [13], [14], and programmable data planes and in-band telemetry give the fabric the visibility to manage it [15], [16], [17].

Several considerations temper adoption. A fabric that is simple to operate through a controller is not simple to understand, and administrators still need fluency in VXLAN EVPN, underlay, and overlay routing, and the quality-of-service configuration that lossless behavior requires; the automation reduces repetitive effort but raises the conceptual floor. Migration from a three-tier network is rarely a clean cut since external Layer-2 and Layer-3 connectivity, data-center interconnect, and policy translation must be handled during coexistence. Reliability at the level of the whole facility, including power and cooling, sits outside the fabric model presented here and must be assessed separately [22]. The analytical model itself carries assumptions worth stating: it treats spine paths as independent, which shared optics or power can violate; it models a homogeneous leaf, whereas production fabrics mix leaf types; and it addresses steady-state capacity rather than transient convergence during failure, which depends on control-plane timers and hardware behavior.

6. Conclusions

Controller-managed spine-and-leaf fabrics have become the default for data centers because they resolve, together, the three problems that hierarchical networks left open: wasted capacity on blocked links, variable path length under east-west traffic, and unmanageable configuration at scale. This paper presented an analytical model that connects the number of spine switches to oversubscription, equal-cost path diversity, bisection bandwidth, and fabric availability, and evaluating that model across representative configurations showed that a single dimensioning decision meets both the performance target and the resilience target, with a non-blocking ratio reached at a determinable spine count. Mapping controller functions to the fabric lifecycle located the operational value of automation in the suppression of configuration drift across build, deploy, and operate stages. The central claim is that deterministic performance at scale is an emergent property of a Clos topology, an EVPN control plane, and controller automation acting in combination and that a design analysis should treat them as one object. Future work will extend the model to transient convergence behavior during spine and leaf failures, incorporate heterogeneous leaf types and realistic traffic matrices, and validate the availability model against longitudinal fabric telemetry.

7. Author Contribution

The sole author conceived the analysis, developed the model, produced the results, and wrote the manuscript.

8. Conflict of Interest

The author declares no conflict of interest. Product names are used for technical identification only and do not imply endorsement.

9. Funding

This research received no external funding.

10. Data Availability Statement

The analytical results reported here are fully reproducible from the equations and parameters stated in the text; no external dataset was used.

REFERENCES

1. C. Clos, "A study of non-blocking switching networks," Bell System Technical Journal, vol. 32, no. 2, pp. 406-424, 1953. Available: https://doi.org/10.1002/j.1538-7305.1953.tb01433.x
2. M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," ACM SIGCOMM Computer Communication Review, vol. 38, no. 4, pp. 63-74, Oct. 2008. Available: https://doi.org/10.1145/1402958.1402967
3. A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri et al., "VL2: A scalable and flexible data center network," in Proc. ACM SIGCOMM 2009 Conf. Data Commun., pp. 51-62, Aug. 2009. Available: https://doi.org/10.1145/1592568.1592576
4. W. Xia, P. Zhao, Y. Wen, and H. Xie, "A survey on data center networking (DCN): Infrastructure and operations," IEEE Communications Surveys & Tutorials, vol. 19, no. 1, pp. 640-656, 1st Quart., 2016. Available: https://doi.org/10.1109/COMST.2016.2626784
5. F. Han, M. Wang, Y. Cui, Q. Li, R. Liang, Y. Liu et al., "Future data center networking: From low latency to deterministic latency," IEEE Network, vol. 36, no. 1, pp. 52-58, Jan. 2022. Available: https://doi.org/10.1109/MNET.102.2000622
6. M. Alizadeh and T. Edsall, "On the data path performance of leaf-spine datacenter fabrics," in Proc. 2013 IEEE 21st Annu. Symp. High-Performance Interconnects, pp. 71-74, Aug. 2013. Available: https://doi.org/10.1109/HOTI.2013.23
7. M. Mahalingam, D. Dutt, K. Duda, P. Agarwal, L. Kreeger, T. Sridhar et al., "Virtual eXtensible Local Area Network (VXLAN): A framework for overlaying virtualized layer 2 networks over layer 3 networks," RFC 7348, IETF, Aug. 2014. Available: https://www.rfc-editor.org/rfc/rfc7348.html
8. A. Sajassi, R. Aggarwal, N. Bitar, A. Isaac, J. Uttaro, J. Drake et al., "BGP MPLS-based Ethernet VPN," RFC 7432, IETF, Mar. 2015. Available: https://doi.org/10.17487/RFC7432
9. D. Kreutz, F. M. V. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," Proceedings of the IEEE, vol. 103, no. 1, pp. 14-76, Jan. 2014. Available: https://doi.org/10.1109/JPROC.2014.2371999
10. A. Leivadeas and M. Falkner, "A survey on intent-based networking," IEEE Communications Surveys & Tutorials, vol. 25, no. 1, pp. 625-655, 1st Quart., 2022. Available: https://doi.org/10.1109/COMST.2022.3215919
11. M. Tavares, R. Valadas, T. Amado, M. Paz, and S. Santos, "Decentralized and dynamic zero-touch provisioning of leaf-spine EVPN data centers," IEEE Access, 2025. Available: https://doi.org/10.1109/ACCESS.2025.3589509
12. J. Zhang, F. R. Yu, S. Wang, T. Huang, Z. Liu, and Y. Liu, "Load balancing in data center networks: A survey," IEEE Communications Surveys & Tutorials, vol. 20, no. 3, pp. 2324-2352, 3rd Quart., 2018. Available: https://doi.org/10.1109/COMST.2018.2816042
13. M. Alizadeh, T. Edsall, S. Dharmapurikar, R. Vaidyanathan, K. Chu, A. Fingerhut et al., "CONGA: Distributed congestion-aware load balancing for datacenters," in Proc. 2014 ACM Conf. SIGCOMM, pp. 503-514, Aug. 2014. Available: https://doi.org/10.1145/2619239.2626316
14. G. Liu, Y. Liu, Q. Meng, B. Wang, K. Chen, and Z. Shen, "Traffic load balancing in data center networks: A comprehensive survey," Computer Science Review, vol. 57, p. 100749, 2025. Available: https://doi.org/10.1016/j.cosrev.2025.100749
15. E. F. Kfoury, J. Crichigno, and E. Bou-Harb, "An exhaustive survey on P4 programmable data plane switches: Taxonomy, applications, challenges, and future trends," IEEE Access, vol. 9, pp. 87094-87155, 2021. Available: https://doi.org/10.1109/ACCESS.2021.3086704
16. W. Quan, Z. Xu, M. Liu, N. Cheng, G. Liu, D. Gao et al., "AI-driven packet forwarding with programmable data plane: A survey," IEEE Communications Surveys & Tutorials, vol. 25, no. 1, pp. 762-790, 1st Quart., 2022. Available: https://doi.org/10.1109/COMST.2022.3217613

17. Z. Xu, Z. Lu, and Z. Zhu, "Information-sensitive in-band network telemetry in P4-based programmable data plane," IEEE/ACM Transactions on Networking, vol. 32, no. 6, pp. 5081-5096, Dec. 2024. Available: https://doi.org/10.1109/TNET.2024.3448244
18. H. Huang, X. Du, T. Li, H. Wang, K. Xu, M. Wang et al., "Re-architecting buffer management in lossless Ethernet," IEEE/ACM Transactions on Networking, vol. 32, no. 6, pp. 4749-4764, Dec. 2024. Available: https://doi.org/10.1109/TNET.2024.3430989
19. Y. Zhang, Q. Meng, C. Hu, H. Zhang, S. Wang, and F. Ren, "ACK-driven congestion control for lossless Ethernet," IEEE/ACM Transactions on Networking, 2025. Available: https://doi.org/10.1109/TON.2025.3542770
20. A. Gangidi, R. Miao, S. Zheng, S. J. Bondu, G. Goes, H. Morsy et al., "RDMA over Ethernet for distributed training at Meta scale," in Proc. ACM SIGCOMM 2024 Conf., pp. 57-70, Aug. 2024. Available: https://doi.org/10.1145/3651890.3672233
21. K. Qian, Y. Xi, J. Cao, J. Gao, Y. Xu, Y. Guan et al., "Alibaba HPN: A data center network for large language model training," in Proc. ACM SIGCOMM 2024 Conf., pp. 691-706, Aug. 2024. Available: https://doi.org/10.1145/3651890.3672265
22. K. M. U. Ahmed, M. H. J. Bollen, and M. Alvarez, "A review of data centers energy consumption and reliability modeling," IEEE Access, vol. 9, pp. 152536-152563, 2021. Available: https://doi.org/10.1109/ACCESS.2021.3125092