A Survey on Internet Multipath Routing and Provisioning

Sandeep Kumar Singh, *Student Member, IEEE*, Tamal Das, and Admela Jukan, *Member, IEEE*

***Abstract*—Utilizing the dormant path diversity through multipath routing in the Internet to reach end users—thereby fulfilling their QoS requirements—is rather logical. While offering better resource utilization, better reliability, and often even much better quality of experience (QoE), multipath routing and provisioning was shown to help network and data center operators achieve traffic engineering in the form of load balancing. In this survey, we first highlight the benefits and basic Internet multipath routing components. We take a top-down approach and review various multipath protocols, from application to link and physical layers, operating at different parts of the Internet. We also describe the mathematical foundations of the multipath operation, as well as highlight the issues and challenges pertaining to reliable data delivery, buffering, and security in deploying multipath provisioning in the Internet. We compare the benefits and drawbacks of these protocols operating at different Internet layers and discuss open issues and challenges.**

***Index Terms*—Multipath routing, traffic engineering, multihoming, QoS, routing, intra/inter-domain.**

I. Introduction

MOST of the layer-based protocols in present Internet establish end-to-end communication over a single path. On the other hand, the emerging online applications—such as search (e.g., Google), gaming, social networking (e.g., Facebook, Twitter), video streaming (e.g., YouTube), real-time applications (e.g., live streaming), on-demand applications—such as Video-on-Demand (VoD), and bandwidth-intensive 4K resolution in ultra high definition television (UHDTV),—pose a new set of constraints on end-to-end delay, throughput, bandwidth, etc and often require routing and provisioning over multiple paths. Thus, multipath routing has emerged as a technology of choice in the Internet, which can fulfill most of the today's application constraints (e.g., throughput, end-to-end delay, bandwidth) as well as network constraints (e.g., load balancing, resource utilization, congestion-control, secure and reliable data delivery). It is being widely used in the form of packet splitting over multiple paths at the network layer (equal-cost multipath in Open Shortest Path first (OSPF) [1] and Intermediate System to Intermediate System (IS–IS) [2]). In addition, the so-called *inverse multiplexing* at the link and optical layers also deploy multiple paths between transmitter and receivers [3]–[5]. The already used methodologies are controlled and mostly operated by the network operators, and in the early days they offered little freedom to end-hosts to make decisions in selecting different routing benefits.

However, recent research proposals on multipath provisioning have opened another dimension in terms of satisfying customers (measured by Quality of Experience or QoE) as well as improving network performance [6], [7]. These proposals emphasize the cross-layers and inter-network (or inter-domain) cooperation to establish end-to-end multiple concurrent paths between hosts, so that the benefit of the path diversity is realized in its true potential. A case in point is Apple's iOS 7 that has implemented Multipath TCP (MPTCP) [7] to transfer data over various mobile interfaces, such as Wi-Fi and 4G LTE [8]. Basically, it uses Wi-Fi interface as a primary TCP connection, and cellular interface as a backup connection to increase resiliency. The concurrent end-to-end paths can exercise a finer control in handling load-balancing and congestion in the network, since end-hosts could adaptively shift load from congested paths to less congested ones. Moreover, bandwidth-intensive applications are more likely to function at higher throughputs by aggregating paths and accordingly splitting the applications' traffic over these concurrent paths.

Over the last decade, multiple parallel (concurrent) lanes (paths) have been used in the metro/regional networks to achieve higher data rates. For example, although 100G Ethernet has been standardized, the mature 10G Ethernet transmission technologies are deployed based on a parallel transmission approach, such as 10 × 10G electrical lane, and 4 × 25G optical lane, to achieve higher data rates [9]. Moreover, in the optical networks, industries have been opting for multipath transmissions using wavelength division multiplexing (WDMs) to increase the data rate rather than achieving high-speed serial transmission over a single lane (as the cost of switching technologies significantly increase with the bit rate). Multipath routing was also shown to facilitate survivability and security features by dispersing data over multiple concurrent paths between end-hosts. Therefore, from a technological perspective, it is evident that multipath routing and provisioning underlies the current Internet and will shape its future.

In this paper, we survey multipath routing and provisioning in the current Internet,[1] from end-user's as well as Internet Service Provider's (ISP's) perspective, and we highlight how the overall network performance can be improved by deploying multipath provisioning across layers and networks. We take a top-down approach and review various multipath protocols,

Manuscript received December 16, 2014; revised June 18, 2015; accepted July 17, 2015. Date of publication July 23, 2015; date of current version November 18, 2015.

The authors are with the Department of Electrical and Computer Engineering, Technische Universität Carolo-Wilhelmina zu Braunschweig, Braunschweig 38106, Germany (e-mail: sandeep.singh@tu-bs.de; t.das@tu-bs.de; a.jukan@tu-bs.de).

Digital Object Identifier 10.1109/COMST.2015.2460222

[1] In this paper, we survey multipathing in wired networks. For multipathing and multicasting in mobile ad-hoc networks, see [10], [11].

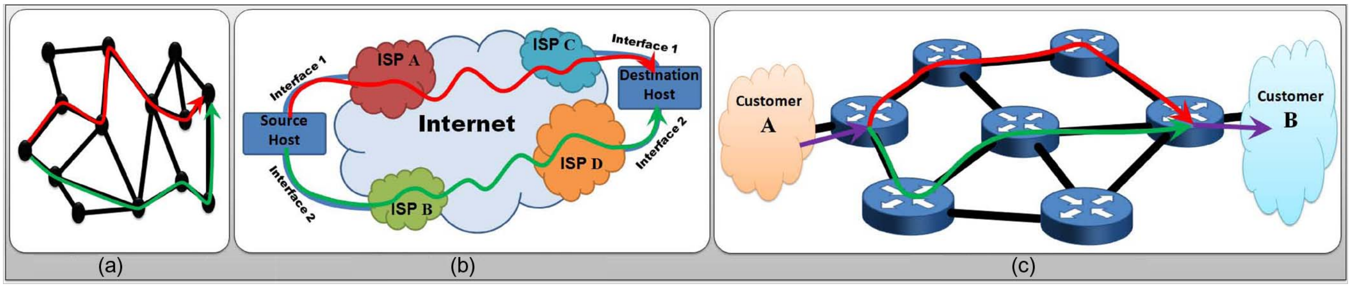

Fig. 1. Illustration of multipath routing. (a) Graph theoretical representation. (b) Realization in the transport layer. (c) Realization in the network layer.

from application to link layers, operating at different parts of the Internet. We also review the mathematical foundations of the multipath operation, as well as highlight the issues and challenges pertaining to reliable data delivery, buffering and security in deploying multipath provisioning in the Internet. For completeness, we also highlight some of the issues and challenges that are hindering the deployment of multipath technologies today. Although, inter-layer cooperation can provide some of the benefits of multipath routing in the intra-domain (i.e., within a network area), inter-network cooperation among ISPs are needed to establish efficient end-to-end data delivery over concurrent paths. More importantly, multiple paths present more challenges than current single path routing in terms of ensuring in-order data delivery. Despite considerable work done in recent times to provision multipath communication in wireline networks—such as the Internet backbone and optical networks—Internet Service Providers (ISPs) are often reluctant to deploy them in real networks due to the capital and operational expenditures incurred in upgrading their networking technologies[2] [12]–[14]. In this survey paper, we address some of these issues and challenges by studying the emerging requirements and mechanisms of multipath provisioning, and review various possible multipath technological solutions.

The rest of this survey paper is organized as follows. Section II highlights benefits of multipath provisioning, such as traffic engineering, QoS routing, load balancing, etc.; basic components of multipath routing—path computation, forwarding and splitting; and multipath routing in the intra-and inter-domain networks. We review existing multipath protocols operating at different layers of the Internet as well as in the lower layers in Section III. Section IV reviews the mathematical analysis and modeling behind some of the proposed multipath algorithms. Section V highlights issues and challenges in deploying multipath provisioning protocols in existing wireline networks, whereas Section VI presents some concluding remarks.

II. Main Features of Multipath Routing

Fig. 1(a) illustrates the path diversity that exist in today's networks, which can be exploited by the network operators to route traffic across diverse paths. However, for the end-users to benefit from the path diversity, the end-to-end concurrent paths should be established by the transport (or application) layer at the end-hosts; only then can end-users take advantage of different ISPs to transmit their data across the Internet. Currently, most of the computer systems can connect to different networks (ISPs), also called *multihoming* technique, by employing any of the these methods: (a) multiple IP addresses for a single network interface, (b) different IP addresses for each interface in a multi-interface system, and, (c) single IP address for multiple network interfaces. In Fig. 1(b), we show two end-to-end paths between hosts that are supported by their dual network interfaces, and carried across multiple networks (ISPs). These multihomed systems thus effectively establish or teardown some of their subflows[3] (of a traffic flow) based on the application requirements or network-state (e.g., congestion). Furthermore, network layer can independently split traffic over similar (e.g., equal-cost) or dissimilar paths at the intermediate routers, which is illustrated in Fig. 1(c).

In this section, we first emphasize the benefits of multipath provisioning, as well as illustrate how multipath routing can replace the prevalent single path routing in current networks, particularly in the Internet backbone. We then discuss the three basic components of multipath routing, namely,

1) a *multipath computation algorithm* to compute multiple paths for a flow,
2) a *multipath forwarding algorithm* to forward packets on diverse paths, and,
3) a *traffic splitting algorithm* to effectively split traffic across multiple paths, subject to requirements such as congestion control, traffic engineering, load balancing, security, resiliency, etc.

We conclude this section with a brief discussion on intra-and inter-domain multipath routing.

A. Benefits of Multipath Routing and Provisioning

Multipath provisioning has emerged as a technology of choice, which can offer numerous advantages as follows.

- *Traffic engineering* (TE), as defined in [15], in the Internet (and valid for most networks) deals with the issue of

[2] Such as softwares and hardwares at the end-hosts (e.g., operating systems, but it is in domain of computer manufacturers), servers, switches, routers as well as middleboxes, such as Network Address translators (NATs), firewalls, proxies.

[3] In this paper, we use the following terms: *flow* to describe an origin to destination (O-D) packet stream of a transport layer connection between two end-hosts; and *subflow* to denote multiple parts of a flow that choose different paths between end-hosts.

performance evaluation and performance optimization of operational IP networks. In other words, TE mechanisms map the traffic flows and network resources of a network in such a way that some of the major objectives, such as reliable communication, higher throughput, minimum delay, congestion control, etc. can be achieved. Utilizing multiple paths between end-hosts is one such TE approach, which can achieve some of these objectives. Apart from establishing multiple end-to-end subflows (of a traffic flow) between end-hosts overmultiple paths, intermediate routers can, in turn, further employ traffic engineering approaches to optimize network utilization. *QoS routing* [16], [17], another aspect of TE, is a routing mechanism, wherein flows are routed based on the availability of network resources (e.g., bandwidth) and QoS requirement (end-to-end delay, delay variance, packet loss rate, etc.) of flows. The exchange of these information (resource availability and QoS requirement) is essential in order to optimize the network to extract maximum performance (e.g., throughput, loss rate, etc.). Concurrent multiple paths can be used to satisfy QoS requirement of multiple flows.

- *Load balancing* and *congestion control* are important aspects of TE and can be achieved by using multipath provisioning. Traffic flows can be distributed over multiple concurrent paths such that all the links are optimally loaded, thereby avoiding network hot-spots. If at all, some of the links/nodes are congested in the network, multipath routing can be efficiently used to shift fraction of the traffic from congested paths to less congested ones. In order to achieve load-balancing, either routers need to disseminate link-load information in the network, or the end-hosts derive the information from the signaling mechanisms.
- *Reliable communication* can be effortlessly realized with the implicit fault tolerance aspect of multipath provisioning. In single path routing, when a path fails (say, from fiber cuts), routing protocols use alternate paths. The application gets interrupted for the transitional time until an alternate path is set up between end-hosts. On the other hand, in the multipath scenario, mostly one or more, but not all, of the concurrent subflows get affected, as they generally use disjoint paths. The lost packets can then be quickly retransmitted over existing non-faulty paths. Hence, the communication remains uninterrupted in the multipath scenario, albeit at lower throughput.
- *Network resource utilization* is another aspect that can be improved by deploying multipath routing. Moreover, higher resource (e.g., bandwidth) utilization often leads to higher network throughput. Applications requiring higher bandwidth than the link capacity cannot be served by single path routing, and are eventually blocked. On the other hand, fractional bandwidths available on the concurrent paths can be aggregated to serve such bandwidth-intensive applications.
- *Security*—While single-path routing is vulnerable to security threats, such as denial-of-service attack (by overloading a particular node/link/path), multipath routing can provide greater security by dispersing data over multiple paths between end-hosts, where each path carries a portion of data between a source-destination pair. Moreover, multipath routing using unpredictable selection of links/paths makes it difficult for an attacker to conduct such attack against any single link/path [18], [19].

B. Finding Multiple Paths

To find multiple paths for a given traffic flow, multipath computation algorithms require a global view of the network topology as well as its resources. In a given scenario, such an algorithm may select paths that are *node-disjoint* (no common nodes except source and destination), *link-disjoint* (no common links) or *non-disjoint* (may have common nodes as well as links). Node/link-disjoint paths improve fault-tolerance and offer more aggregate bandwidth than non-disjoint paths, since a *bottleneck* node/link failure for non-disjoint paths can severely impact the performance of multiple paths [20]. However, non-disjoint paths are easy to discover due to no constraint on common nodes and links.

The performance of a multipath computation algorithm depends on the *number* and *quality* of paths selected. Optimal number of paths for a flow is desirable for various reasons, such as, (a) to reduce overhead in establishing, maintaining and tearing down multiple paths, (b) to reduce complexity of a multipath routing scheme, which considerably increases with the number of paths, and, (c) to accommodate constraints on the number of explicitly routed paths, such as label-switched paths (LSPs) in Multi-Protocol Label Switching (MPLS) [21], that can be set up between a pair of nodes [22]. It is thus a trade-off between network congestion and the number of parallel paths.

The quality of parallel paths depends on availability of network resources at intermediate routers, characteristics of paths and the QoS requirements. The availability of network resources, which reflects the congestion in the network, can be obtained through periodic exchange of (link-load) information between routers. While computing paths, non-disjoint (or *bottleneck*) links should be avoided. The common characteristics of paths are bandwidth, latency, cost, and reliability. QoS requirements of flows are often expressed as combination of bandwidth requirements and delay bounds. Thus, different applications have different QoS requirements (such as high throughput or low latency), and the service providers ask the infrastructure providers to provision their services, while satisfying their QoS requirements. In QoS-based routing [16], only a subset of paths that altogether satisfy the QoS requirements are selected.

A number of multipath computing algorithms have been proposed in the literature to find node-disjoint paths [23]–[27] as well as link-disjoint paths [25], [28], [29] between the end-hosts in the network. However, these algorithms do not consider QoS parameters, or link-congestion status as a metric while computing paths. To meet certain objectives like congestion minimization, QoS requirements, etc. path computing algorithms must have the knowledge of current status of network congestion, utilization, etc., but they are difficult to measure and generally not known to the intermediate routers. However, there are proposed algorithms to find disjoint paths [22], [30]–[33], wherein QoS requirements of flows and resource availability information are periodically exchanged among network nodes, and therefore routing decisions are made based on those information. Authors in [34]–[37] consider differential path delays as a constraint to find multiple paths in the optical transport networks.

Path computation algorithms can be broadly classified into two categories, namely, protocols based on distance-vector (DV; based on Bellman-Ford algorithm) and link-state (LS; based on Dijkshtra's algorithm)—the latter being predominant in today's Internet [38]. These protocols require periodic (or intermittent) exchange of connectivity information between routers (after any link/node failure). In DV routing, be it single path (Routing Information Protocol [39], DUAL [40]) or multipath (MPATH [41], [42]), each router only informs its neighbors of its distance to all nodes in the network, based on which, the routers compute the shortest path to each node. On the other hand, in LS routing, such as OSPF and IS–IS, each router broadcasts the status of each of its adjacent links, based on which, the routers derive the network topology and compute the shortest path to each node. LS protocols are more suitable to implement multipath computation algorithms, because each node has complete map of the network [43]. Moreover, LS protocols are more suited for multipath routing due to (a) shorter route convergence time in case of failure of links/nodes, (b) ability to compute more complicated routes than DV protocols, and, (c) their (OSPF and IS–IS) predominance in today's Internet [43], [44].

Numerous works investigated multipath routing based on a Path Computation Element (PCE)-based architecture [45]–[50]. [45] proposed a routing algorithm to find two optimal disjoint (primary and backup) QoS paths across multiple IP/MPLS domains based on a PCE-based architecture, which works fully decoupled from the Border Gateway Protocol (BGP). [46] proposed a PCE-based source routing scheme to calculate QoS paths (with a given domain sequence) in multiple IP/MPLS domains (or autonomous systems), however, without guaranteeing end-to-end disjoint paths. On the other hand, [47] finds end-to-end shortest disjoint survivable paths in forward and backward directions by using PCE-based signaling in multi-domain networks. [48] investigated and quantified the benefits of multipath routing for distributed data-intensive application with high bandwidth requirements and multidomain reach; based on the PCE-architecture, it proposed *segmental* and *end-to-end* multipath routing. [49] presented a backward-compatible inter-domain multipath routing framework without significant changes to the existing inter-domain routing protocols. It banks on the network abstraction technologies to compose multiple virtual routing planes and offers two variants of multipath mechanisms. [50] proposed multipath extensions in PCE, covering both protocol extensions as well as implementation. While all related prior works implemented multipath routing using single-path routing capabilities of each PCE, this one proposed fundamental modifications to the `PCReq` and `PCRep` messages of the PCE Protocol (PCEP) [51] to enable multipath routing per-PCE.

OpenFlow offers group abstraction to represent a set of ports as a single entity for forwarding packets, thereby enabling multipath provisioning [52]. Openflow Link-layer Multipath Switching (OLiMPS) aims to improve efficient, manageable use of large networks by optimizing dataflow mapping in complex multipath topologies [53]. Two Internet-scale SDN-based testbeds to evaluate and promote MPTCP for experimenters and early adopters was launched based on OpenFlow support of GÉANT and PlanetLab Europe [54].

Multipath communications have been highly recommended for datacenter networks, overriding the spanning tree protocol, for better network utilization prospects. [55] proposed SDN-based rack-to-rack multipath switching in datacenter networks for load balancing purposes, whereas [56], [57] proposed SDN-based load balancing for fat-tree datacenter networks with multipath support. [58] proposed and analyzed a multipath-multicast file transfer scheme to reduce transmission times for one-to-many file transfers in OpenFlow networks.

C. Forwarding on Multiple Paths

Once intermediate routers and end-hosts compute the connectivity (path) information, the subsequent question that arises is: how to *forward* packets along these multiple paths? Forwarding is a method which maps incoming packets to outgoing links. In today's IP network, packets are forwarded along a shortest path by the intermediate routers based on their destination addresses.

Some of the existing shortest-path forwarding mechanisms that can be utilized in multipath packet forwarding are as follows.

- *Destination-based (hop-by-hop) forwarding*: On arrival of an IP packet, a router performs a "longest prefix matching" of the packet's destination IP address and its forwarding table entries, and accordingly forwards the packet on the applicable outgoing link. Multihomed end-hosts' packets (having different IP addresses) can be easily forwarded over multiple paths. However, hop-by-hop forwarding does not offer a very flexible method for mapping of packets over multiple paths [59].
- In the *source-based (explicit) routing* approach, either a path between a source-destination pair can be specified in the IP packet's header (using IP options), or it can be explicitly specified through a group of labels (MPLS-based explicit routing). Though explicit routing simplifies the role of intermediate routers and reduces the forwarding table size, it does not efficiently utilize the bandwidth in comparison to destination-based forwarding.
- *Hashing* is a well-known approach to map an incoming packets to an outgoing link [60]. The router performs a hash (e.g., CRC16) on the packet's header, and the packet is then forwarded on the outgoing link that matches (maps) the outcome of the hash function (i.e., key). On performing a flow-based hash function, all packets of an *n*-tuple TCP flow are forwarded to the same outgoing link, since all packets of that flow have the same *n*-tuple

header information (e.g., for $n = 5$, 5-tuple TCP header has: IP addresses, port numbers and protocol type).

D. Traffic Splitting Along Multiple Paths

Once a set of paths for a given flow are determined and a forwarding technique is selected, the source node can begin sending data to the destination along the specified paths. But an important issue that still remains is: how to *distribute* traffic over multiple paths? Broadly, traffic splitting algorithms can be classified into *even splitting* (such as Equal-Cost Multi-Path (ECMP) [1], [61], Sridharan *et al.* [62]) and *uneven splitting* (such as OSPF Optimized Multipath (OSPF-OMP) [63], Table-based Hashing with Reassignments (THR) [64] and TeXCP [65]). While the former distribute incoming traffic *uniformly* over the next-hop links, the latter distribute traffic *non-uniformly* based on factors such as congestion, bandwidth availability, etc. Uneven traffic splitting algorithms achieve better performance at the cost of increased overhead in periodically updating link load information between all routers.

In general, traffic splitting algorithms can be classified as follows [59]:

- *Round-robin* is the classic mechanism to evenly split a traffic flow over multiple paths. It splits at the granularity of packets, and hence a desired packet splitting ratio can be accurately achieved. However, this leads to excessive packet reordering at the destination, as the packets of a TCP flow experience different delays over different paths, which further degrades the TCP throughput performance. TCP considers the out-of-order packet delivery as an indicator of packet loss, and hence, the TCP source reduces the packet transmission rate. Therefore, round-robin forwarding works well only for the multiple paths having similar path delay.
- In a *per-flow traffic splitting* technique, each router maintains per-active flow state information, i.e., it maps the packets of a flow to its pre-assigned path. On the other hand, a new flow has the flexibility to route dynamically to any of the available paths. However, forwarding table size increases with the increase in the number of flows. Therefore, this technique is suitable for small-sized networks only.
- *Burst (or flowlet) based traffic splitting* functions at the granularity of burst (i.e., a group of packets), and forwards the (two) successive packets of a flow over multiple paths in such a way that the inter-packet spacing is larger than the delay difference between the (two) paths [66]. This ensures in-order packet delivery at the destination. However, flowlet-based traffic splitting needs the differential path delay to be known at the diverging point (router), which requires an additional signaling mechanism and overhead.
- *Per-packet traffic splitting* technique, on the other hand, can (un)evenly split the incoming packets of a stream (flow) at finer granularity, and splitting is performed by the intermediate routers based on the destination of packets as well as the network-state information (e.g., link-load, congestion). Therefore, it can achieve better load balancing in the network than the per-flow splitting. However, packets need to be reordered at the destination.

The performance (end-to-end delay) of per-packet and per-connection (or per-flow) based multipath source routing have been analytically studied for a two-node network [67]. It has been showed that per-packet allocation exhibits better performance than the per-flow allocation. However, packet-reordering is a major concern in a per-packet multipath routing for connection-oriented transport services. It has been widely researched [68]–[76] over varieties of networks using single-path as well as multipath transport protocols, and it has been concluded that excessive packet reordering can lead to overall performance degradation (e.g., low throughput). One solution is to select only similar (equal-hops/costs) paths in order to mitigate the different path delays. Other method is to use buffer at the receiver to re-sequence packets. Therefore, per-packet multipath routing needs to carefully handle the reordering problem to leverage multiple concurrent paths.

E. Intra/Inter-Domain Multipath Routing

In an intra-domain scenario, one of the easiest way to employ multipath routing is to use source routing, where end-hosts use the topological information to compute multiple paths to a destination and then explicitly embed a particular next hop address into a packet's header, as required. However, this approach lacks scalability, inefficiently utilizes available resources (due to the lack of freedom in routing [77]), has serious security concerns [78], and the most practical hindrance in its deployment is that the current IP routers do not support it. Currently, intra-domain routing protocols, such as OSPF and IS–IS, allow equal-cost multipath traffic splitting, and EIGRP (Enhanced Interior Gateway Routing Protocol), a distance-vector protocol, allows unequal cost multipath routing [79]. Another approach is to use multiple logical topologies, using OSPF [76], and IS–IS [80], on top of a physical topology. Multi-Topology (MT) routing allows routers to assign different weights to each link so that the link is treated by the traffic as a set of independent links, and based on the application QoS requirement, traffic can be spread over these multi-weight links [81], [82]. However, packet forwarding is not so easy in MT routing due to the multiple instances of logical topologies, and packets need to be routed over the same logical topology to avoid loops [77].

Inter-domain routing, on the other hand, mainly uses an exterior gateway protocol BGP [83] in the Internet to exchange the reachability information among ASes. Currently, a standard BGP algorithm selects a single "best route" for a destination, and advertises it to its peers. It does not advertise multiple paths, even though there might be existing some alternate paths between multihomed ASes (i.e., an AS that connects to multiple ASes). There are several proposals [49], [84]–[87] to extend BGP (by changing path selection rules) for multipath operations, as well as some vendors like Cisco [88] and Juniper [89] have come up with routers that support multiple BGP paths. An extension to BGP uses `ADD-PATH` capability to

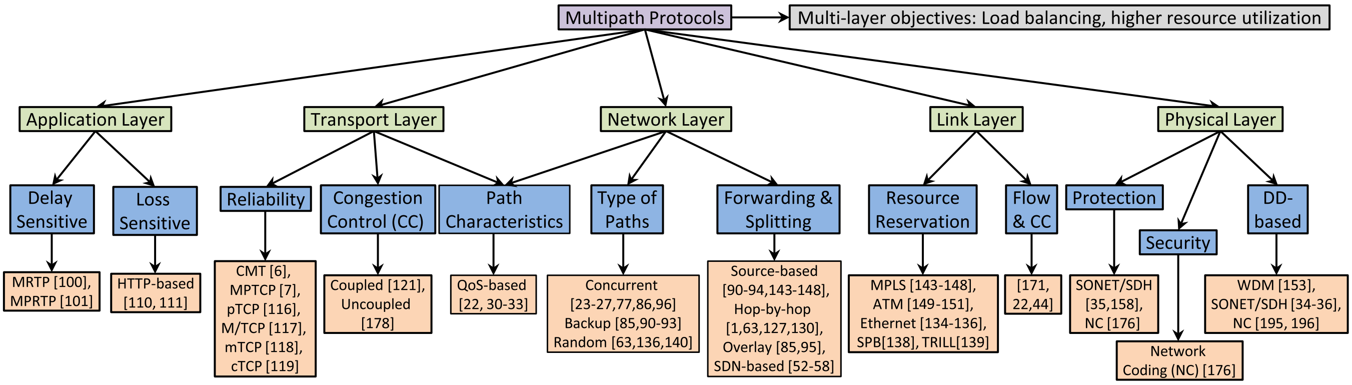
Multipath Protocols
Multi-layer objectives: Load balancing, higher resource utilization
Application Layer
Transport Layer
Network Layer
Link Layer
Physical Layer
Delay Sensitive
Loss Sensitive
Reliability
Congestion Control (CC)
Path Characteristics
Type of Paths
Forwarding & Splitting
Resource Reservation
Flow & CC
Protection
Security
DD-based
MRTP [100], MPRTP [101]
HTTP-based [110, 111]
CMT [6], MPTCP [7], pTCP [116], M/TCP [117], mTCP [118], cTCP [119]
Coupled [121], Uncoupled [178]
QoS-based [22, 30-33]
Concurrent [23-27,77,86,96] Backup [85,90-93] Random [63,136,140]
Source-based [90-94,143-148], Hop-by-hop [1,63,127,130], Overlay [85,95], SDN-based [52-58]
MPLS [143-148], ATM [149-151], Ethernet [134-136], SPB[138], TRILL[139]
[171, 22,44]
SONET/SDH [35,158], NC [176]
Network Coding (NC) [176]
WDM [153], SONET/SDH [34-36], NC [195, 196]

Fig. 2. Classification of multipath features based on layers.

advertise, identify and add multiple paths to a destination [84]. Another possible approach could be source routing [90]–[94], which can establish an end-to-end multiple path, but unlike in intra-domain, inter-domain routing is managed by multiple ISPs. Hence, the inflexible source routing gives little control to the transit ASes over transit traffic, and that might lead to the blocking of the transit traffic, if consent is not given by the transit ASes. Resilient Overlay Networks (RONs) [95], on the other hand, do not require consent from the transit ASes; they form a virtual topology on top of the existing Internet. RON can find multiple routes among RON-enabled nodes (i.e., overlay nodes at different ASes) in case of any problem in the underlining Internet paths. To forward packet over alternate paths, IP-in-IP tunnel mechanism is used (i.e., IP packets are encapsulated within a new IP header). RON, however, does not have control over paths between nodes and involves extra overheads. MIRO (Multi-path Inter-domain ROuting) is another overlay approach that creates demand-based alternate paths (thereby reducing overheads) between two ASes by negotiation [85]. Thus, it gives better control over paths.

Recently, a backward compatible BGP multipath solution, called BGP eXtended Multipath (BGP-XM), has been proposed in [96], which selects multiple routes received from different ASes and advertises them using an aggregated attribute, such as `AS_PATH`, into a single update message without violating usual BGP functionality. While the non-cooperative game-theoretical-based models have been applied in several engineering applications, such as routing, resource allocation and pricing [97], some proposals (see [98], [99]) focus on inter-domain multipath routing. [98] presents a non-cooperative multipath peering framework for coordination between two providers while considering both routing and congestion costs. They show that the coordination policies can decrease the routing cost around 10%, improve route stability, and avoid peering link congestion. On the other hand, [99] utilizes the multi-AS cooperation to ensure edge-to-edge load-balancing and resiliency.

III. From Application Layer to Link Layer

In this section, we review various layer-based protocols that have been proposed to provision multiple paths between end hosts. The important question that arises with the networking protocols is— *at which layer can the maximum benefit of the multipath provisioning be obtained with least overhead and complexity*? Therefore, we evaluate the merits and drawbacks of these protocols running at different layers. Fig. 2 classifies the multipath features at different layers in terms of delay/loss sensitivity, reliability and congestion control, path characteristics and forwarding/splitting, resource reservation and flow-control, and protection and security.

A. Application Layer

Most of the current applications rely on transport or lower (i.e., network or link) layers to deliver content using multipath provisioning. As a result, applications have less control over path selection, data forwarding, etc., as these prerogatives come under the domain of lower layers. However, many real-time media applications (e.g., live streaming, VoD) cannot use multiple paths because while splitting traffic over multiple paths, a tighter delay constraint is not being enforced by the transport/network layer protocols. Hence, for real-time multimedia applications, real-time multipath transport protocols—such as MRTP (Multi-flow Real-Time Protocol) [100] and MPRTP (Multi-Path Real-Time Protocol) [101]—are additionally required to meet the SLA (service-level agreement) and QoS requirements. Moreover, deploying several "similar" paths for a stream would increase the data delivery rate, since they pool cumulative capacity of the multiple paths. A multipath real-time protocol (RTP), based on the single path RTP, must support end-to-end data delivery with real-time characteristics [102], such as delay bound, constant playout rate, etc. Furthermore, an RTP stream is generally carried over UDP/IP (though TCP/IP can also be used), which has no inherent congestion control. Therefore, the receiver requires substantial buffer to compensate the different path latencies encountered by the packets of an RTP stream in order to play at constant bit rate (CBR).

MRTP [100] provides an end-to-end multipath transport service to real-time data over multiple paths using multiple flows. MRTP has been developed and tested for ad-hoc wireless networks, and is claimed to be applicable for the Internet [100]. MPRTP splits a single media stream into multiple subflows, which are then routed over multiple paths [103]. It uses RTP control protocol (RTCP) to monitor the end-to-end data delivery and control information, such as synchronization information, loss, jitter, etc. The de-jitter buffer at the receiver compensates the differential path delays and reorders the out-of-order packets. Although some specific congestion control mechanisms for real-time applications are yet to be tested and integrated with MPRTP [104], the scheduler uses per-path control information and round-trip time (RTT) to judiciously shift traffic among *congested*, *mildly congested*, and *non-congested* paths.

BitTorrent [105], a peer-to-peer (P2P) file sharing protocol, can be considered as an example of application layer multipath protocol, where a user downloads different chunks of a file from various peers across diverse locations.[4] As the order of arrival of these chunks at the receiver is irrelevant, the only thing that matters is to successfully retrieve all chunks of the file. Another example of multipath at application layer could be to use multiple parallel transport (for example, TCP) connections for video streaming, web requests and responses in the Content Delivery Networks (CDNs) [106], [107]. Authors in [108], [109] propose application-level stripping schemes to improve throughput, and creates multiple parallel TCP sockets (connections) for an application. However, these approaches (multiple TCP connections) follow a single path, and do not leverage the dormant path diversity in the network. On the other hand, recent works [110], [111] on multipathing in CDNs have focused on using HTTP (hyper text transfer protocol) due to its capability to cross over middleboxes, such as NAT routers and firewalls, seamlessly. [110] proposed a multipath adaptive HTTP streaming scheme that pre-encodes multimedia content in different bit rates, and splits it into segments of the same duration. A software agent in the user terminal requests these segments from several HTTP servers in parallel using HTTP byte range request. Implementation shows a greater throughput and improved users' QoE. [111] proposes a similar idea based on HTTP range request feature to fetch different content chunks from different servers, however socket interface needs modification at the receiver end.

B. Transport Layer

Although multipath provisioning can be implemented at different layers, transport layer proves to be the best place to provision an end-to-end connection of an application over disjoint paths. Currently, the end-to-end characteristics of a single-path are estimated mainly at the transport layer. Hence, multipath provisioning at this layer can utilize these characteristics for each path to achieve numerous advantages over single-path mechanisms. Moreover, most of the hosts are nowadays supported by more than one network interface, such as, Ethernet, WiFi and cellular. Hence, the transport layer can use them to establish multiple paths.

The idea to use concurrent paths for end-to-end connection between hosts at transport layer is several decades old [112]–[114]. However, one of the first proposals was suggested in 1995 [115]. Thereafter, many other multipath transport protocols were proposed [6], [7], [116]–[120]. In [116], Hsieh *et al.* propose *pTCP* (parallel TCP)— a TCP-based protocol, as an end-to-end transport protocol to strip the data over multiple paths in wireless domains. Another TCP-based protocol for multipath operation is *M/TCP* [117], which uses "multi-route" identity option in TCP header to send data over multiple paths, and maintains "information lists" of transmitted segments over each route to overcome the difficulties in handling acknowledgments and retransmissions. *mTCP* [118]—an extension to TCP—also strips data over multiple end-to-end paths, and deploy a shared bottleneck detection mechanism to identify congested links and suppress some of the mTCP subflows in order to not utilize an unfairly larger share of bandwidth than other single-path TCP flows. mTCP selects a number of candidate paths for a connection using underlying resilient overlay networks (RON) [95]. Hence, it cannot be used on an arbitrary IP network [6]. *cTCP* (concurrent TCP) is another approach of load balancing at transport layer that can be employed to establish multiple paths for a connection between multihomed end-hosts [119]. Concurrent Multipath Transfer (CMT) [6]—an extension to SCTP—utilizes the multihoming feature to concurrently transfer data over multiple independent paths. Use of a single sequence space (and not separate for each path) for an *association* across all paths have been proposed and argued to be beneficial in handling congestion control, loss detection and recovery of data traffic. However, sharing of a single sequence space across paths leaves gaps in it, and therefore packets might be blocked by the middleboxes, such as Network Address Translators (NATs) in present Internet [12].

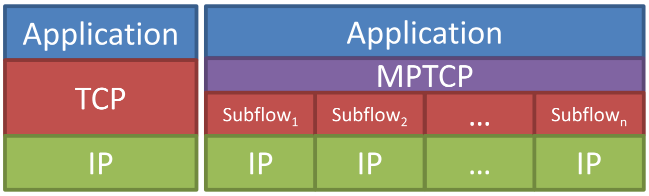

Fig. 3. A comparison of TCP and MPTCP stacks.

Recently, an Internet Engineering Task Force (IETF) working group standardized the multipath protocol for the transport layer. They extended the regular TCP to *Multipath TCP* (MPTCP), which allows a transport connection to be split into multiple "subflows" simultaneously, and operate across multiple paths [7]. These subflows use either different IP addresses or different port numbers with a same IP address (of multihomed end-hosts) to exchange packets belonging to a single MPTCP connection over concurrent (possibly disjoint) paths. MPTCP appears like a regular TCP connection to the application layer. However, the network layer treats each MPTCP subflow as a regular TCP flow, carrying extra information in TCP option field. Fig. 3 shows the regular TCP and MPTCP protocol stacks.

MPTCP was designed to [12], [121]:

- *Improve throughput*—the performance of MPTCP should be at least as good as TCP.
- "*Do no harm*"—MPTCP must work in all scenarios, where TCP currently works, and it should not take up more capacity than a single-path flow per path chosen.
- *Balance congestion*—MPTCP must shift traffic from congested paths to less congested ones by efficient handling of subflows.

[4] It is an example of multi-source to a single destination peering.

MPTCP starts like a regular TCP, and then sets up additional subflows if there exists multiple paths between a multi-homed (or, multi-addressed) endhosts. A subflow is essentially a TCP flow of an MPTCP connection, which can be setup like a TCP flow with some extra signaling mechanism. All MPTCP signals are carried in the optional header field (40 bytes) of TCP. To achieve high throughput, MPTCP manages a separate *congestion window* for each subflow, and the size of the windows is varied based on the congestion level measured on each subflow-path [121]. A congestion window indicates the amount of data a sender can transmit without being acknowledged on a connection. Furthermore, the congestion windows on different subflows are coupled at the connection-level to ensure that the data can be shifted from congested paths to less congested ones, and the aggregate bandwidth used by the MPTCP connection should not be more than a single TCP connection.

Finally, Multipath TCP has also been applied in Data Centers (DCs) [122]. To this end, MPTCP has been compared against PacketScatter (spreads packets of a flow over multiple paths), regular TCP and two TCP variants (Uncoupled-TCP and Equally-Weighted TCP) applied on simulation-based Fat-Tree topology [122], [123]. Although PacketScatter achieves the highest throughput and best load balancing, in-order delivery remains an issue, especially when the lengths of the shortest and longest paths differ a lot. MPTCP outperforms regular TCP and the above-mentioned TCP variants.

C. Network Layer

In Section II-B, we discussed multipath computation mechanisms and associated algorithms based on distance-vector or link-state approaches, different forwarding techniques, and various traffic splitting approaches. Some of the well-known routing protocols in the Internet (e.g., OSPF, IS–IS in the intra-domain, and BGP in the inter-domain) originally did not support multipath operations. However, currently, with the extensions of these protocols, it is possible for intermediate routers to split traffic flows over multiple next hops (or outgoing links) [61], [63]. Here, we review some of the well-received proposals that enable network layer[5] protocols to perform multipath operation.

Various routing protocols, such as OSPF version 2 [1] and ISIS-OMP [124], allow *equal-cost multipath* (ECMP) routes to be selected by the routers, and then split packets of different traffic flows over multiple shortest paths of equal cost [61]. ECMP-enabled OSPF protocol attempts to *evenly* distribute traffic across the multiple next hops to achieve load balancing at the network layer. Traffic can be split based on per-packet or per-flow, and ECMP mostly uses per-flow splitting of traffic among multiple paths. Therefore, packets of a traffic flow always follow the same route. However, the load balancing at flow-level (i.e., coarse granularity), is achieved by uniform distribution (using round-robin mechanism) of traffic flows over multiple paths. This uniform distribution of flows and the use of only equal-cost shortest paths limits the congestion-control and load balancing capability of the ECMP [44]. To forward a packet over next hops, the router can choose any of the following methods [18], [63].

- *Hash-Threshold* uses the 16-bit Cyclic Redundancy Check (CRC-16) to perform a hash over the packet header fields that define a flow. The key, generated by the hash function, falls in any of the hash spaces that are assigned to next hop candidates. The hop which satisfies the mapping (key to hash spaces of hops), is chosen as the next hop for that packet.
- *Modulo-N* selects a next-hop among *N* hops by performing modulo-*N* hash over the packet header.
- *Highest Random Weight* (HRW) obtains a separate weight for *each* next-hop by hashing the packet header fields by its seed (a random number which is assigned to each next-hops separately). The next-hop which obtains the highest weight is selected for the packet. This method is less disruptive (due to addition or removal of next-hops) and more expensive than Hash-Threshold.

ECMP only allows equal-cost shortest paths to be selected for an even distribution of traffic flows, and it does not attempt to make changes in link-costs for load balancing. However, if there is only a single shortest path available for some traffic flows, then it does not distribute those traffic flows on the alternate paths. Moreover, an even distribution of traffic is not always optimal and preferable for load balancing [63].

There have been some attempts to change link weights according to flow requirements (e.g., delay, jitter, loss) and routing policies to achieve load balancing in the networks (multi-topology routing [81], [82], ARPANET [125]), but these load-sensitive routing methods suffer from the route instability caused by routing oscillations (i.e., traffic switches between different paths).

OSPF Optimized Multipath (OSPF-OMP) is an extension to OSPF routing protocol, which is designed to allow uneven traffic splitting across the equal cost paths [63]. Generally, routers do not exchange (link or path) loading information among themselves, therefore, uneven splitting to achieve load balancing becomes nearly impossible. On the other hand, to divide the traffic *unevenly* over multiple next-hops, OSPF-OMP floods the link loading information within an OSPF area, and path loading information regarding an adjacent area is flooded by an Area Border Router (ABR).

Routing deflection [126] is a tag-based routing method in which end-systems use tags to select multiple (not necessarily shortest) paths. In this method, routers "deflect" packets to some of its neighbors based on the deflection rules. The end-systems then construct available end-to-end multiple paths, and use tags to inform intermediate routers of which paths they want to use for their packets.

A recently proposed network layer solution *Locator/ Identifier Separation Protocol* (LISP) [127] divides IP functionality into two parts for localization and identification of a device, by assigning two IP addresses for mainly scalability reason. The localization IP addresses, usually belong to border

[5] The objective of network layer multipath protocols are to route data from source to destination by computing paths, and then forwarding and splitting the traffic flows over desired paths.

routers, are called Routing LOCators (RLOCs), and identification IP addresses of endpoints are named as Endpoint Identifiers (EIDs). LISP border routers advertise their RLOCs so that an EID can reach them, and creates a tunnel to forward endpoints' packets between RLOCs routers by prepending a destination RLOC IP address. LISP border routers can split a flow over multiple paths using different destination RLOCs addresses.

Apart from the ISPs, *Datacenter (DC) operators* often employ time-tested Internet routing protocols within DCs. On comparing intra-DC traffic with the Internet traffic, recent DC traffic characterization studies, however, found that ISP routing protocols are ineffective in DC environment to achieve high link utilization, leading to congested core links (of tree-based DCs) [128], [129]. Moreover, routing protocols in DCs should be able to proactively (or reactively) decongest the bottleneck (core) links by offloading some traffic onto under-utilized links. Thus, TE in DCs require customized routing protocols, to fully exploit its architectural design and better utilize network resources to enhance load balancing and reduce packet loss.

Equal cost multiple path (ECMP) forwarding—most prevalent multipath routing used in today's DC—evenly spreads the flows over the next hops using hash functions. Fixed Hashing (mapping), performed on packet header, forwards all packets of a TCP flow onto a single path, thereby ensuring in-order packet delivery at destination (otherwise, excessive packet reordering may deteriorate TCP performance). Although per-flow-based ECMP route packets of a flow onto a single path, many simultaneous TCP connections between end-hosts (having same pair of IP addresses, but different pair of port numbers) are routed on different paths in order to achieve load balancing. However, two or more large, long-lived flows can collide on their hash, and end up on the same output port of a switch, creating an avoidable bottleneck [123]. Moreover, ECMP fails to exploit increased path diversity that can be provided by traffic engineering techniques through the assignment of non-uniform link weights to optimize network resource usage [130].

MicroTE is a TE mechanism that adapts to microscopic (per-second) variations in network traffic patterns [131]. It uses a central controller to aggregate and create a global view of network conditions and traffic demands, and computes routes with the objective of minimizing the maximum link utilization. MicroTE uses weighted ECMP to stripe unpredictable traffic (at the flow-level) across the residual capacity along candidate paths for the traffic. However, MicroTE only slightly outperforms ECMP and it exhibits 1% to 15% deviation from the optimum. The major shortcoming of MicroTE is its assumption of traffic matrix predictability, which is highly unlikely in a DC environment.

Penalizing Exponential Flow-spliTing (PEFT) [132] is a link-state routing protocol, which forwards packets based on link weights on a hop-by-hop basis. Unlike OSPF, PEFT-enabled switches split traffic unevenly (exponentially) over all possible paths (shorter and longer), but longer paths are penalized based on total link weights along the paths. In the modified PEFT [130], an online traffic measurement module is incorporated to compute the optimal set of link weights, and to monitor link utilization of individual links. The modified PEFT achieves better load balancing and link utilization as compared to ECMP on tree-based topologies (Canonical Tree, Fat Tree and Interlink Tree) [130].

D. Link Layer

Link layer mainly reserves the link bandwidths for the data packets, and switches them from input ports (or interfaces) to output ports (or interfaces). Here, we review some of the link layer protocols which are used in Ethernet, MPLS and ATM networks. Although MPLS works between layers 2 and 3, and ATM is a combination of both these layers, we choose to describe them in this subsection, and not in Section III-C. Currently, Ethernet is widely used in local area networks (LANs) and metropolitan area networks (MANs). However, in today's *virtual private networks* (VPN), MPLS is preferred. ATM, on the other hand, was originally developed to meet the requirements of voice, video and data traffic in broadband ISDN (Integrated services digital Network).

Multipath in Ethernet Networks: Ethernet technology has been used in LANs and MANs since several decades. In traditional LANs, Spanning Tree Protocol (STP) runs on bridges and switches. STP is a simple link layer protocol to provide a "loop free" switching path between all pairs of nodes in a LAN. The Ethernet learning mechanism helps every node to create a tree topology, such that it can forward packets to other nodes. Ethernet is a preferred technology for MAN and clustered networks, but there are certain shortcomings, which needs to be eliminated. The root cause of performance bottleneck and lower fault-tolerance threshold is the single spanning tree (SST) switching used in these networks. Apart from the under-utilized links (SST blocks many links to construct a tree) and scaling problems, it takes about 30 to 60 seconds to re-converge routing, which is unacceptable in metro and cluster networks.

Multiple Spanning Tree Protocol (MSTP) was proposed to overcome shortcomings of STP [133]. Multiple switching paths between a pair of end-hosts can be provisioned using multiple spanning tree instances. Multiple paths provide an automatic scope for load balancing and fault-tolerance. However, the use of MST in networks is non-trivial, as the switches need a mechanism to distinguish between various spanning trees for effectively maintaining the MAC address forwarding database [134]. Viking, proposed for MAN and cluster networks, uses Virtual LAN (VLAN) technology with the multiple spanning tree approach using IEEE 802.1s standard [134], [135]. Viking leverages on this facility to come up with load-balanced switching paths, which can be explicitly selected by specifying the VLAN tags associated with the corresponding spanning trees. This selection can be configured by the end-hosts, instead of the network switches. Therefore, different VLAN identifiers are tagged to the packets of different flows traversing on multiple paths (formed by MST). However, major disadvantage with Viking is that it uses a centralized algorithm to compute multiple paths, and relies on external tools to obtain the network topology. Finally, SPAIN (Smart Path Assignment In Networks) provides multipath forwarding using inexpensive, COTS (commodity off-the-shelf) Ethernet switches over arbitrary DC topologies [136]. It pre-computes a set of paths that exploit the redundancy in a given network topology, then merges these paths into a set of trees; each tree is mapped as a separate VLAN onto the physical Ethernet. However, SPAIN requires some software modification at end-hosts to choose pre-installed paths.

Recently, a backward compatible IEEE 802.1aq standard, called *Shortest Path Bridging* (SPB) [137], [138], was approved by IETF, which uses Shortest Path Tree (SPT) as an alternative to STP and related protocols like Rapid STP and MSTP. SPB allows all equal cost paths to be active, by distinguishing them using multiple VLANs, in the Ethernet networks, and end stations are expected to use ECMP to simultaneously forward packets over these paths. It encapsulates customer frame with an Ethernet header, either by MAC-in-MAC or VLAN identifier-based method, route it in the SPB domain using IS–IS protocol. *Transparent Interconnection of Lots of Links* (TRILL) [139] is another link-layer forwarding protocol, which operates within Ethernet broadcast domain, and uses IS–IS routing to compute shortest paths in the network. TRILL switches, also called Routing Bridges (RBridges), exchange IS–IS Hello frames to discover each other. The edge RBridges encapsulate Ethernet frames with a TRILL header, route it in the TRILL domain using IS–IS, and finally decapsulate the Ethernet frames. Multipathing is supported using hash-based per-flow (instead of per-packet) ECMP routing in order to avoid packet reordering. Both TRILL and SPB (supported by vendors like Cisco, Huawei) make network highly resilient and efficient, and they are promising for data centers.

Valiant Load Balancing (VLB) is adopted for the VL2 data center—a randomized traffic spreading technique over a virtual layer-2 infrastructure, which essentially guarantees equal-spread load-balancing in a mesh network [140]. Using VLB, each server randomly picks a path (core switch) for each flow. VL2 uses per-flow VLB (effectively equivalent to ECMP), which forwards packets of a flow on the same path to avoid packet reordering. *Hedera* is a dynamic, centralized flow scheduling technique (a layer-2 protocol), which collects flow information (demand) from constituent switches regularly, computes "good" paths and installs them on the switches [141]. It then reschedules "large" flows (those exceeding 10% of the host-NIC bandwidth) from highly utilized links to under-utilized ones, while "mice" flows are routed by ECMP. Hedera achieves higher aggregate network utilization as compared to ECMP, when applied on multi-rooted tree topologies (e.g. Fat-Tree). However, the major drawback with Hedera is that it has to run every 0.5 seconds, while utilization improvement over ECMP is only 1% to 5% [122], [142].

Multipath in MPLS Networks: In MPLS [21], paths (also known as label switched paths or LSPs) are identified by the edge label switched routers (LSRs), and they encapsulate the layer 3 packets into their header that contains a fixed size label, class of service, and other information, such as flags, time-to-live (TTL), etc. Core LSRs then forward the MPLS packets based on the matching of the label and the next-hop in their forwarding table, and a new label replaces the old one, which helps it scale in the Internet. LSPs are similar to virtual circuits; therefore, packets of a flow can be assigned multiple labels to route them along multiple paths, but the packets that have to follow the same route must carry the same label.

MATE is an MPLS-based adaptive scheme, in which *probe* packets are used to measure *one-way* LSP statistics, such as packet delay and packet loss [143]. These measurements are then used to decide on when and how to shift traffic among the LSPs. Basically, load-distribution is performed by distributing traffic equally among fixed number of *bins*, and then, each bin is mapped to the corresponding LSPs according to MATE algorithm. [144] proposed a constraint-based (hop-count) multiple paths computing algorithm for MPLS network to minimize the maximum link utilization. It further proposes a traffic splitting algorithm to obtain a load splitting ratio for the pre-computed paths [145]. [146] proposed to use multiple LSPs as back-up routes in case of failures of primary routes in the MPLS networks. [147] developed heuristics as well as linear programming-based traffic bifurcation method that minimizes the maximum link utilization in the MPLS networks. Self-Protecting Multipath (SPM) [148] is a protection mechanism for MPLS networks, wherein traffic is transmitted over multiple paths with path failure notification and a load redistribution function controlled by the source node.

Other Link Layer Multipath Concepts: Asynchronous Transfer Mode (ATM) networks were proposed in the 1990s for its features such as bit rate flexibility, statistical multiplexing capabilities and QoS guarantees. In ATM, early proposals related to multipathing used a number of end-to end virtual circuits between a source and destination [149], [150]. [149] proposed a burst-level bandwidth reservation scheme for the multi-path routing, while [150] established parallel communication between hosts in ATM networks. In [151], Pareta *et al.* established multiple virtual paths (VPs) for carrying connectionless traffic of source-destination pairs. Packets are probabilistically split along VPs based on the relative utilization of the links of the VPs in ATM networks.

E. Multipath at Physical Layer-Fiber Networks

While a few decades ago, the bandwidth offered by a single-mode fiber was more than sufficient for any application, the growth of Internet and emergence of bandwidth-intensive, high-performance applications pushed the network resources (including bandwidth) to the limits. As the (circuit-and packet-switched) WDM and the inverse multiplexing technologies (e.g., virtual concatenation [157]) in the optical domain are maturing, the effective utilization of bandwidth (offered by multiple wavelengths in a multi-mode fiber) is realizable in practice today [158]. Recently, many research proposals have come up with parallel transmission in WDM networks to utilize the aggregated residual bandwidth of multiple wavelengths across fibers. Chen *et al.* considered multipath provisioning in WDM networks to satisfy the high bandwidth demands of applications that cannot be met by any single wavelength [153]. It formulated an integer linear programming (ILP) based approach to find multiple paths with differential path delay and number of fiber delay lines (FDLs) as constraints. Huang *et al.* proposed heuristics to compute multiple paths and split traffic as a mechanism to provide survivability in SONET/SDH, which can be used on top of WDM networks [35].

SONET/SDH is a circuit-switched technology that was originally developed to carry a constant bit rate voice traffic in the optical transport networks [158]. Although data frames are carried over a single path, an alternate path is used to carry the same data under 1 + 1 protection scheme. This results in carrying more than 100% overhead, and hence the inefficient utilization of bandwidth, which is not desirable in today's networks. Therefore, in the next-generation SONET/SDH, efforts have been made to support different data traffic, and integrate them with the voice traffic in SONET/SDH frames. *Data over SONET/SDH* (DoS) has been developed as a transport mechanism that uses the three techniques, namely generic framing procedure (GFP) [159], virtual concatenation (VCAT) [157], and link-capacity-adjustment scheme (LCAS) [160]. GFP is used to map data traffic into the SONET/SDH frame; VCAT is an inverse multiplexing technique, which allows a connection to be split over multiple paths; and LCAS can dynamically allocate bandwidth by adding/removing the paths associated to the data traffic. DoS accommodates the different types of data packets/frames, such as IP, Ethernet, MPLS, ATM, into SONET/SDH frames such that the DoS frames can be routed over multiple (two for SONET/SDH ring) paths in optical transport networks [5].

Without VCAT, a STS-48c ($\approx$ 2.5 Gbps) SONET container is needed to carry a Gigabit Ethernet connection, which results in about 60% wastage of bandwidth [161]. On the other hand, if VCAT is used to split a connection over virtual containers then seven STS-3c ($\approx$ 155.5 Mbps each) would be sufficient to carry a gigabit Ethernet connection, which results in only about 10% wastage of bandwidth. There are several research proposals [34]–[36], [161]–[163] that consider different constraints, such as differential delays, overhead, bandwidth utilization, while establishing multiple paths for an Ethernet over SONET/SDH (EoS) connection. [162] proposed PESO, a reliable EoS transport mechanism, which guarantees a certain limit on degradation of a connection in case of a link/node failure; routes data with low bandwidth overhead as compared to SONET/SDH; and reduces the failure recovery time (in the order of 50 ms) as compared to Ethernet STP.

Multipath provisioning algorithms of WDM networks can easily be extended for spectrum allocation and survivability in elastic optical networks (EONs) that divides the traditional fixgrid spectrum spacing (e.g., 50 GHz) into several mini-or flexi-grids [154]–[156]. Moreover, with the advent of sliceable bandwidth-variable transponders, also called multiflow optical transponders, it is now possible to serve various low capacity (e.g., 40G) and high capacity (400G) demands using multiple parallel optical flows [164].

F. Cross Layer Multipathing

Although the implementation of cross-layer interaction is complex in nature due to the association of different features of control and data planes at different layers, it can be beneficial for efficient multipath routing and provisioning between end-hosts. The higher layers (Application and Transport) have a poor knowledge of the underlying network-wide path diversity between end-hosts, therefore lower layers (Network and below) must share useful knowledge with higher layers to boost performance gain offered by multipath-solution [169]. Cross-layer cooperation is also useful in handling flow and congestion control in the networks [170]. [169] focuses on transport and network layer cross-signaling that allows MPTCP endpoints to establish additional subflows using a Locator/Identifier Separation Protocol (LISP) in the cloud network. LISP-capable border routers allow splitting of MPTCP subflows over different Internet paths, if not end-to-end then at least along a segment of the Internet path. A cross-layer Augmented MPTCP (A-MPTCP) achieves better performance (file transfer time), with just one additional subflows, than original MPTCP and legacy TCP. [171] proposed a holistic cross-layer multipath communication architecture (however, without an implementation) for cloud networking, which utilizes MPTCP, LISP and TRILL at different layers, respectively to achieve a higher throughput. Many recent works have used layer 2 switching concepts, particularly OpenFlow, in conjunction with MPTCP. Researchers at CalTech and CERN employed OpenFlow controller and MPTCP to achieve a transfer rate of 339 Gbps in 2012 [172]–[174]. Furthermore, optical layer has also been used in conjunction with layer 2 (see previous subsection). For example, 100 Gigabit Ethernet is mapped to 10 optical subcarriers and subsequently routed over disjoint fiber-level paths [175]. Furthermore, it has been shown recently in [176] that fault-tolerant design using network coding is feasible for high-speed Ethernet over optical networks.

Finally, we summarize the multipath protocols operating at different layers in Table I. It also highlights topology on which they are efficient and mainly used, type of paths used, decision making layers (which enforce multiple paths to be used), whether routes are computed as centralized or distributed, the main objective for deploying these routing protocols and if they are available or not in the market.

IV. Modeling and Analysis of Multipath Routing

In this section, we first compare the multiple paths computing algorithms, path-disjointedness, and their objectives in Table II. We then review the network performance optimization models, and subsequently we review the stochastic models of multipath routing.

A. Complexity of Multipath Routing Algorithms

The very first step in routing on single or multiple path(s) is to disseminate topology information to all the nodes (routers) in the network. The nodes then can forward the traffic along multiple paths to each destination. However, as compared to single path routing (in OSPF, single shortest path computation complexity is $O(N \log N)$, where N is the number of nodes in the network), multiple path computing algorithms are generally more complex (for example, K-shortest paths computational complexity in a network with N nodes is $O(KN + N \log N)$) [26]. Internet pose another challenge, as the end-to-end paths more often fall in multiple domains (networks). The inter-domain routing protocols (mainly BGP) and their multipath extensions would need to exchange connectivity information

TABLE I
CLASSIFICATION OF MULTIPATH ROUTING PROTOCOLS BASED ON DIFFERENT PARAMETERS

Layers[1]	Protocols	Topology	Type of Paths[2]	Centralized/[3] Distributed	Objective(s)[4]	Market[5] Availability
L5	MPRTP [101], MRTP [100]	Arbitrary	Concurrent	Distributed	Delay-sensitive	NA
	HTTP-based [110], [111]	Arbitrary, CDNs	Concurrent	Distributed	Loss-sensitive	NA
L4	MPTCP [7]	Arbitrary, DCs	Concurrent	Distributed	LB, Reliability	A
	MPTCP-CC [121]	Arbitrary	Concurrent	Distributed	CC, LB	NA
	CMT [6]	Arbitrary	Concurrent	Distributed	LB, Reliability	NA
L3	OSPF-OMP [63]	Arbitrary	Equal-cost paths	Distributed	Load balancing	A
	ECMP [1]	Arbitrary, DCs	Equal-cost paths	Distributed	Load balancing	WA
	Routing deflection [126]	Arbitrary	Multiple	Centralized	Load balancing	NA
	LISP [127]	Arbitrary	Equal-cost paths	Distributed	Load balancing	A
	PEFT [130], [132]	Arbitrary, DCs	Multiple	Distributed	Link utilization	NA
	MicroTE [131]	DC (Fat-tree)	(Un)equal-cost paths	Centralized	Link utilization	NA
	Packet-Scatter	Arbitrary	Multiple (Random)	Distributed	Load balancing	A
	Multipath BGP-based [77], [86], [96]	Arbitrary	Concurrent	Distributed	Link utilization, LB	A
	Inter-domain-based [85], [90]–[93]	Arbitrary	Backup	Distributed	Reliability	NA
L2	MPLS-based [143]–[148]	Arbitrary	Pre-computed	Centralized	LB, Protection	NA
	ATM-based [149]–[151]	Arbitrary	Pre-computed	Centralized	Load balancing	NA
	Viking [134]	Arbitrary	Pre-computed	Centralized	Load balancing	NA
	SPAIN [136]	Arbitrary	Multiple (Random)	Centralized	Link utilization	NA
	TRILL [139]	Arbitrary	Equal-cost paths	Distributed	Load balancing	A
	SPB [138]	Arbitrary	Equal-cost paths	Distributed	Load balancing	A
	Hedera [141]	DCs	Multiple	Centralized	Load balancing	NA
	VLB [140]	DC (Fat-tree)	Equal-cost paths (Random)	Distributed	Load balancing	NA
	Portland [152]	DC (Fat-tree)	Equal-cost paths	Distributed	VM migration	NA
L1	SONET/SDH-based [34]–[36]	Arbitrary (Optical)	Pre-computed	Centralized	DD, Reliability, Overhead	NA
	WDM-based [153]	Arbitrary (Optical)	Pre-computed	Centralized	BP, RU, LB	NA
	EON-based [154]–[156]	Arbitrary (Optical)	Pre-computed	Centralized	Survivability, RU, LB	NA

[1] L5: Application layer, L4: Transport layer, L3: Network layer, L2: Data-link layer, L1: Physical (Optical) layer.
[2] Multiple: shortest & non-shortest paths. Random: Packets are randomly split on one of many available paths. Pre-computed: Pre-setup paths (e.g., LSP).
[3] Centralized and distributed are with respect to path setup mechanism.
[4] LB: Load balancing, CC: Congestion control VM: Virtual machine, DD: Differential delay, BP: Blocking probability, RU: Resource utilization.
[5] NA: Not Available, A: Available or Announced, WA: Widely Available.

TABLE II
COMPARISON OF VARIOUS MULTIPATH ALGORITHMS IN TERMS OF SELECTION OF PATHS AND TRAFFIC SPLITTING METHODS

Multipath Routing Algorithms[1]	Complexity[2]		Criteria[3]	
	Path Computation	Packet Splitting	Path Selection	Packet Splitting
Suurballe [23]	$O(n^2 \log n)$	–	n pairs of ND paths*	–
S-T algorithm [28]	$O(m \log_{(1+m/n)} n)$	–	n pairs of LD paths*	–
Yen's algorithm [165]	$O(KN(m + N \log N))$	–	*Loopless* K-shortest paths**	–
Eppstein [26]	$O(m + KN + N \log N)$	–	K-shortest paths between a pair of nodes	–
OSPF-ECMP*[1]	–	$O(K)$	Non-D, Equal-Cost SPs	Even, per-flow, Hash-based
OSPF-OMP*[124]	–	–	Equal-Cost SPs	Uneven (based on LPLI)
Nelakuditi *et al.**[22]	$O(KN^2)$	$O(K)$	Non-D,*** Path metric (ARLB)	Equalized blocking probability
LDM [166]	$O(N^2)$	$O(K)$	Non-D, minimum hop-count	Probabalistic distribution
MATE [143]	–	$O(K)$ or $O(K^2)$*	All possible paths	Equal delay derivatives, packet loss
DASM*[167]	**	–	Link-costs	Length of the Labels
MPATH**[41]	$O(N^2 + N^2 \log N)$	–	Loop-free, unequal-cost paths	–

[1] * Link-State (LS), and ** Distance-vector (DV) routing algorithms.
[2] K, N, m: number of paths (or LSPs in MPLS networks) per O-D pair, nodes, and edges (links) in the network, respectively.
* If hash-based splitting is used, complexity is $O(K)$, otherwise if *gradient projection algorithm* is used, complexity is $O(K^2)$ [169].
** After a single link failure or link-cost increase, *time complexity* is $O(x)$, where x is the number of affected routers; and, after a single link addition or link-cost reduction, it is $O(d)$, where d is the network diameter (i.e., longest shortest path in hops between any two routers).
[3] Paths: ND (Node Disjoint), LD (Link disjoint), Non-D (Non-Disjoint), *LPLI (Link and Path load information), ARLB (Average Residual Link Bandwidth).
* between a source and n sinks. ** K-shortest paths between a node to all other nodes. *** disjoint w.r.t. bottleneck links.

among ASes. In Table II, we list some of the multipath routing algorithms, their complexity, criteria for selection of multiple paths as well as splitting.

B. Optimization Models

As discussed in Section II-A, multipath presents several benefits, such as congestion control by load balancing, higher link (or network) utilization by efficiently rerouting of traffic, etc. To achieve these objectives, some of the proposed routing algorithms either use linear programming (LP) optimization models or heuristics. Although LP-based optimization provides optimal solutions, their computational complexities rarely scale. On the other hand, heuristics provide sub-optimal solutions, but can scale well in large networks. The fundamental objectives of these algorithms are to compute paths and adaptively distribute traffic among those paths. Heuristic-based ECMP has been proposed for balancing the load across multiple shortest paths, using a simple round-robin distribution [61]. OSPF-OMP, on the other hand, unevenly splits traffic among paths based on a flow distribution algorithm [63]. Nelakuditi *et al.* present a heuristic scheme to proportionally split traffic over widest disjoint paths [22]. These widest paths are selected such that they are disjoint with respect to bottleneck links, which eventually help minimize congestion. Banner *et al.* [44] presented a theoretical model to prove the intractability of minimizing network congestion under path constraints (i.e., restriction on link-length and number of paths per destination), and an approximate solution is also proposed.

In recent decades, congestion control was widely investigated on varieties of networks. Multipath routing has been considered as good candidate to deal with congestion control. However, the current TCP congestion control mechanism (for example, additive-increase/multiplicative-decrease) need to be changed to take full advantage of path diversity [177]. This is due to the fact that it operates as an uncoupled mechanism, which independently tries to handle the congestion on each path between end-hosts; and it does not coordinate among subflows (of a connection) at the transport layer. To evaluate the congestion control mechanism for multipath TCP over multiple paths, Raiciu *et al.* [178] proposed four algorithms: namely, Fully coupled, Linked Increases, Uncoupled TCPs, and RTT Compensator. These algorithms couple the congestion windows of various subflows in different ways. MPTCP, as described in Section III-B, uses coupled congestion control mechanism to efficiently shift load from highly congested subflows (traversing on those congested paths) to lesser ones [121].

The congestion control mechanisms, however, have been mostly studied along with other network performance measures, such as link and network utilization, since they are mutually related. Most of the optimization models try to minimize the maximum link utilization. The basic optimization model to deal with utilization function was proposed by Kelly *et al.* to control the demand into the network [179]. It maximizes the following utility function.

$$max \sum_{\forall k} U_k(x^k),$$

where, x^k is the input rate of the k^{th} user (or traffic source), and $U_k(\cdot)$ is the associated utility function. The solution of this optimization problem is tractable only if the objective function is strictly concave, and if the feasible region is convex [179]. It has been noted in [147] that if the single-path routing is imposed using this optimization, it then complicates the optimization through non-convexity. However, if multiple routes are allowed, then the problem is well-behaved. [177] proposes a similar optimization-based congestion control scheme for multipath TCP and extends the proof of single-path stability condition to multipath congestion-control stability (i.e., this algorithm is able to find a split and stability *if the flows can be split among the available routes such that the total load on each link is less than the link capacity*). [180] finds a sufficient condition for the local stability of end-to-end algorithms for joint (multipath) routing. The stability condition as observed by authors is: *the responsiveness of each route is restricted by the round-trip time of that route alone, and not by the round-trip times of other routes*.

Wang *et al.* proposed an LP-based approach to minimize congestion of the most utilized links in MPLS-based networks, considering multipath routing [147]. It emphasizes that minimizing the maximum link utilization leaves more space for future traffic growth, provided traffic grows (scales up) in proportion to the current traffic pattern. Another similar LP (min-max problem) was also proposed to utilize multipaths, with an additional constraint on number of hop-counts [145]. MATE—a distributed adaptive load-balancing algorithm for MPLS networks—controls a flow's path between ingress and egress nodes, and utilizes multiple paths in order to avoid congestion in the MPLS networks [143]. It effectively shifts traffic from high to less loaded LSPs (Label Switched Paths). A Distributed Adaptive Traffic Engineering (DATE) algorithm was proposed to jointly optimize the dual goals (maximize throughput, minimize congestion) of end users and network operators and quickly react to avoid bottlenecks [181]. DATE is a multipath routing protocol, where the edge routers split traffic for each source-destination pair over multiple paths. A distributed online traffic engineering approach (TeXCP) was also presented based on multipath routing [65]. Like most traffic engineering schemes, it also aims to minimize link utilization. The TeXCP agent uses light-weight explicit feedback from the core routers to discover path utilization, and adaptively moves traffic from over-utilized paths to under-utilized ones.

C. Stochastic Models

Early works involve multipath routing combined with resource reservation modeled using continuous-time Markov chains in the context of ATM Private Network-Network Interface (PNNI) standard [182] and Broadband Integrated Services Digital Networks (B-ISDN) [183]. [184] modeled (using continuous-time Markov chain), analyzed and compared multipath algorithms with single path algorithms that might be persistent, i.e., retry after a failure. In terms of network throughput, multipath routing performed slightly better than single path routing with no retries, and slightly worse with one or two retries. In terms of connection establishment time, multipath routing performed significantly better than single path routing.

[69] investigates the extent of reordering introduced in a packet stream being forwarded along different paths, using a packet reordering metric known as Reorder Density [185], while [186] presents a comparative analysis of various packet reordering metrics—such as Reorder Density (RD), Reorder Buffer-occupancy Density (RBD), Reorder extend and n-Reordering—with respect to *essential* and *desirable* attributes. Essential attributes include capturing reordering in a sequence, low sensitivity to packet loss and duplication and metric's usefulness to evaluate behavior and performance of a network, while desirable attributes include simplicity, informativeness, evaluation complexity, robustness and extensibility to cascaded networks.

[187] proposed two traffic congestion control techniques for multipath communications, namely, *flow assignment* (to optimize splitting of traffic across multiple paths, thereby reducing end-to-end path delays) and *packet scheduling* (to reduce packet resequencing delay and resequencing buffer occupancy). It assumed a multiple-node $M/M/1$ tandem network with a fixed delay line as a path model, and Gaussian distributed end-to-end path delays.

[188] analyzed the effect of fixed delay in conjunction with queueing and resequencing delays on the optimal (with respect to minimal total end-to-end delay) traffic distribution on multiple disjoint paths in high-speed networks, and concluded that the optimal splitting may heavily depend on the difference in the fixed delays on the two paths, as well as the bandwidth available on the paths. It models the queueing delay by an equivalent exponential service rate, and the fixed portion of the delay by a delay line. [189] proposed a packet resequencing model for high-speed networks to estimate the packet (resequencing/total) delay and the resequencing buffer occupancy distributions based on queueing theory concepts.

V. Open Issues and Challenges

Although multipath provisioning presents several benefits, these come with greater complexity, and there are some issues that need to be addressed in the near future so that its potential can be fully realized in the operational networks. We highlight some of the issues and challenges here.

A. Routing Stability

Dynamic routing algorithms should respond quickly to topological and traffic changes in the network to achieve better network performance. However, the frequent updates of routing messages can lead to network instability due to routing oscillation, and hence performance degradation [180], [190]. The route oscillation (where traffic switches between different paths) problem can occur in intra/inter-domain, single-path or multipath routing protocols [191]. In original ARPANET [125], delay was used as a routing metric, and routing oscillation was observed under heavy load condition. Multi-topology routing also suffers from it, as it tries to change the link weights [81], [82]. Dynamic multipath routing poses greater difficulty, since traffic has to be split on multiple paths, without compromising stability. Recent works, such as MATE [143], TeXCP [65], have been successful in providing stable multipath operation.

B. Routing Complexity, Overhead and Scalability

In multipath routing, network topological information need to be disseminated to all nodes so that they can utilize the path diversity, while computing the multiple paths (currently, link-state protocols do this job in the Internet). These operations present increased routing complexity (than single-path) and cost additional overhead. Moreover, intermediate routers have to store additional routes per destination, which increase the routing table size. More importantly, as the routing complexity increases, its scalability reduces. Therefore, optimal number of paths per source-destination pair is required to reduce complexity, memory use and computation power [22]. Furthermore, [192] shows that multipath routing benefits decrease or performs closer to single path routing when network size grows, typically with the number of nodes $N > 25$ and the number of links $L \ll D$ (i.e., the number of demand pairs).

C. Differential Delay, Packet Reordering and Buffering

As multipath routing comprises of multiple paths, an implicit issue is the difference in delay encountered by packets of a flow routed on the parallel paths between a given source and destination—termed as the *differential delay* issue. This, in turn, causes the packets to arrive out-of-order at the destination, and for to ensure in-order delivery, they have to be resequenced before feeding to the application. Moreover, per-packet splitting at the intermediate routers can further increase the reordering, and as discussed in Section II-D, excess packet reordering can lead to performance degradation for connection-oriented transport services. One solution is to store the packets in *resequencing* buffer at the destination. However, possibly this will lead to large *destination buffering* requirements, which is often expensive. An alternative, however, is to distribute the buffering requirement along the path over intermediate nodes, instead of only at the destination, which is termed as *distributed buffering* [193]. Solutions have been proposed for adaptive multipath routing to optimize flow assignment and packet ordering at the source, based on network parameters to minimize packet reordering at the destination [187]. [194] proposed optimization models and heuristics to solve the routing and distributed buffering (or delay compensation) in the context of inverse multiplexed optical transport network (OTN) architectures. *Network coding* has also been proposed to minimize the destination buffer size for handling differential delay [195], [196].

D. Cross-Layer and Inter-Network Cooperation

After reviewing multipath techniques operating at different layers, it is clear that network (or link) layer can route (switch) data traffic across multiple paths, however, to measure the path characteristics (e.g., RTT, available bandwidths, congestion-state) and to satisfy the QoS requirements of traffic flows (or TE in general), multipath provisioning should be originated and controlled at the higher layers (Transport and above). However, multipath provisioning is more complex at higher layers than at the lower layers due to monitoring of path characteristics by signaling mechanisms, and thereafter actions taken based on the measurements. Therefore, to optimize the network resources and to extract the best possible performances, multipath provisioning has to be deployed across layers. But, today's Internet infrastructure is controlled by a large group of network operators. Therefore, inter-network cooperation is necessary to establish multiple end-to-end paths. As described in Section II-E, ISPs are reluctant to deploy inter-domain multipath protocols due to their business policies. Therefore, a new business model has to evolve that could benefit ISPs as well as end-users.

VI. Conclusion

In this paper, we surveyed multipath routing and provisioning in the current Internet. We highlighted the benefits of utilizing the path diversity for the traffic engineering and beyond. We highlighted the issues and challenges pertaining to stability, scalability, packet reordering, buffering, inter-layer and inter-domain cooperation in multipath provisioning. This survey showed that multipath routing is advantageous from both network operator's and end-user's perspectives. Furthermore, multipathing can be better utilized at higher layers (transport and application) to establish end-to-end concurrent paths and exercise a finer control in handling load balancing, congestion control and fault tolerance in the network. Also, during failure, multipath routing can enable redundant (in $1 + 1$ protection scheme) or non-redundant concurrent paths to handle network survivability.

This survey on benefits, issues and challenges of multipath routing and provisioning showcased few important lessons for various stakeholders (as well as layers) involved in delivering QoE/QoS. So far, layer-4 (transport) has been best suited for a practical implementation of multipathing, since end-to-end path characteristics are best known to end-hosts (at layer-4). Similarly, while multipath flow/packet scheduling can be done at any layer, path characteristics can be best used by transport layer to establish end-to-end concurrent paths. We expect to see more research in cross-layer and inter-domain signaling that can help in operation, administration and maintenance of concurrent paths in all layers. In conclusion, multipath routing and provisioning continue to underlie the current Internet and will shape its future and further innovation.

References

[1] J. Moy, "OSPF version 2," IETF, Fremont, CA, USA, RFC 2328, 1998.

[2] R. W. Callon, "Use of OSI IS–IS for routing in TCP/IP and dual environments," IETF, Fremont, CA, USA, RFC 1195, 1990.

[3] E. Gustafsson and G. Karlsson, "A literature survey on traffic dispersion," *IEEE Netw.*, vol. 11, no. 2, pp. 28–36, Mar./Apr. 1997.

[4] C. M. Corbalis, R. M. Moley, S. K. Sathe, and U. Schmidt, "Asynchronous transfer mode communication in inverse multiplexing over multiple communication links," U.S. Patent 5 617 417, Apr. 1, 1997.

[5] D. Cavendish, K. Murakami, S.-H. Yun, O. Matsuda, and M. Nishihara, "New transport services for next-generation SONET/SDH systems," *IEEE Commun. Mag.*, vol. 40, no. 5, pp. 80–87, May 2002.

[6] J. R. Iyengar, P. D. Amer, and R. Stewart, "Concurrent multipath transfer using SCTP multihoming over independent end-to-end paths," *IEEE/ACM Trans. Netw.*, vol. 14, no. 5, pp. 951–964, Oct. 2006.

[7] M. Handley, A. Ford, C. Raiciu, and O. Bonaventure, "TCP extensions for multipath operation with multiple addresses," IETF, Fremont, CA, USA, RFC 6824, 2013.

[8] iOS: Multipath TCP Support in iOS 7. [Online]. Available: http://support.apple.com/en-us/HT201373

[9] P. Winzer, "Beyond 100G Ethernet," *IEEE Commun. Mag.*, vol. 48, no. 7, pp. 26–30, Jul. 2010.

[10] M. Tarique, K. E. Tepe, S. Adibi, and S. Erfani, "Survey of multipath routing protocols for mobile ad hoc networks," *J. Netw. Comput. Appl.*, vol. 32, no. 6, pp. 1125–1143, Nov. 2009.

[11] L. Junhai, Y. Danxia, X. Liu, and F. Mingyu, "A survey of multicast routing protocols for mobile ad-hoc networks," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 1, pp. 78–91, 1st Quart. 2009.

[12] C. Raiciu *et al.*, "How hard can it be? Designing and implementing a deployable multipath TCP," in *Proc. NSDI*, 2012, vol. 12, pp. 29–29.

[13] M. Yannuzzi, X. Masip-Bruin, and O. Bonaventure, "Open issues in interdomain routing: A survey," *IEEE Netw.*, vol. 19, no. 6, pp. 49–56, Nov./Dec. 2005.

[14] M. Chamania and A. Jukan, "A survey of inter-domain peering and provisioning solutions for the next generation optical networks," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 1, pp. 33–51, 1st Quart. 2009.

[15] D. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao, "Overview and principles of Internet traffic engineering," IETF, Fremont, CA, USA, RFC 3272, May 2002.

[16] E. Crawley, H. Sandick, R. Nair, and B. Rajagopalan, "A framework for QoS-based routing in the Internet," IETF, Fremont, CA, USA, 1998.

[17] X. Xiao and L. M. Ni, "Internet QoS: A big picture," *IEEE Netw.*, vol. 13, no. 2, pp. 8–18, Mar./Apr. 1999.

[18] D. Thaler and C. Hopps, "Multipath issues in unicast and multicast next-hop selection," IETF, Fremont, CA, USA, RFC 2991, Nov. 2000.

[19] P. P. Lee, V. Misra, and D. Rubenstein, "Distributed algorithms for secure multipath routing," in *Proc. IEEE INFOCOM*, 2005, vol. 3, pp. 1952–1963.

[20] S. Mueller, R. P. Tsang, and D. Ghosal, "Multipath routing in mobile ad hoc networks: Issues and challenges," in *Performance Tools and Applications to Networked Systems*, M. C. Calzarossa and E. Gelenbe, Eds. Berlin, Germany: Springer-Verlag, 2004, pp. 209–234.

[21] E. Rosen, A. Viswanathan, and R. Callon, "Multiprotocol Label Switching Architecture," IETF, Fremont, CA, USA, RFC 3031, 2001.

[22] S. Nelakuditi and Z.-L. Zhang, "On selection of paths for multipath routing," in *Proc. IWQoS*, 2001, pp. 170–184.

[23] J. Suurballe, "Disjoint paths in a network," *Networks*, vol. 4, no. 2, pp. 125–145, 1974.

[24] K. Ishida, Y. Kakuda, and T. Kikuno, "A routing protocol for finding two node-disjoint paths in computer networks," in *Proc. IEEE Int. Conf. Netw. Protocols*, 1995, pp. 340–347.

[25] R. G. Ogier, V. Rutenburg, and N. Shacham, "Distributed algorithms for computing shortest pairs of disjoint paths," *IEEE Trans. Inf. Theory*, vol. 39, no. 2, pp. 443–455, Mar. 1993.

[26] D. Eppstein, "Finding the k shortest paths," *SIAM J. Comput.*, vol. 28, no. 2, pp. 652–673, 1998.

[27] R. Bhandari, "Optimal physical diversity algorithms and survivable networks," in *Proc. 2nd IEEE Symp. Comput. Commun.*, 1997, pp. 433–441.

[28] J. W. Suurballe and R. E. Tarjan, "A quick method for finding shortest pairs of disjoint paths," *Networks*, vol. 14, no. 2, pp. 325–336, Summer 1984.

[29] D. Sidhu, R. Nair, and S. Abdallah, "Finding disjoint paths in networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 21, no. 4, pp. 43–51, Sep. 1991.

[30] A. Orda and G. Apostolopoulos, "QoS routing mechanisms and OSPF extensions," in *Proc. IEEE GLOBECOM*, 1999, pp. 1903–1908.

[31] Q. Ma and P. Steenkiste, "On path selection for traffic with bandwidth guarantees," in *Proc. Int. Conf. Netw. Protocols*, 1997, pp. 191–202.

[32] Z. Wang and J. Crowcroft, "Quality-of-service routing for supporting multimedia applications," *IEEE J. Sel. Areas Commun.*, vol. 14, no. 7, pp. 1228–1234, Sep. 1996.

[33] Y. Guo, F. Kuipers, and P. Van Mieghem, "Link-disjoint paths for reliable QoS routing," *Int. J. Commun. Syst.*, vol. 16, no. 9, pp. 779–798, 2003.

[34] S. Ahuja, M. Krunz, and T. Korkmaz, "Optimal path selection for minimizing the differential delay in Ethernet-over-SONET," *Comput. Netw.*, vol. 50, no. 13, pp. 2349–2363, Sep. 2006.

[35] S. Huang, C. U. Martel, and B. Mukherjee, "Survivable multipath provisioning with differential delay constraint in telecom mesh networks," *IEEE/ACM Trans. Netw.*, vol. 19, no. 3, pp. 657–669, Jun. 2011.

[36] A. Srivastava, S. Acharya, M. Alicherry, B. Gupta, and P. Risbood, "Differential delay aware routing for Ethernet over SONET/SDH," in *Proc. IEEE INFOCOM*, 2005, vol. 2, pp. 1117–1127.

[37] S. Rai, O. Deshpande, C. Ou, C. U. Martel, and B. Mukherjee, "Reliable multipath provisioning for high-capacity backbone mesh networks," *IEEE/ACM Trans. Netw.*, vol. 15, no. 4, pp. 803–812, Aug. 2007.

[38] A. S. Tanenbaum,*Computer Networks*, 4th, ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2003.

[39] R. Atkinson and M. Fanto, "RIPv2 cryptographic authentication," IETF, Fremont, CA, USA, RFC 4822 (Proposed Standard), Feb. 2007. [Online]. Available: http://www.ietf.org/rfc/rfc4822.txt

[40] J. J. Garcia-Lunes-Aceves, "Loop-free routing using diffusing computations," *IEEE/ACM Trans. Netw.*, vol. 1, no. 1, pp. 130–141, Feb. 1993.

[41] S. Vutukury and J. Garcia-Luna-Aceves, "MPATH: A loop-free multipath routing algorithm," *Microprocess. Microsyst.*, vol. 24, no. 6, pp. 319–327, Oct. 2000.

[42] J. Chen, P. Druschel, and D. Subramanian, "An efficient multipath forwarding method," in *Proc. IEEE INFOCOM*, 1998, vol. 3, pp. 1418–1425.

[43] C. Huitema, *Routing in the Internet*. Upper Saddle River, NJ, USA: Prentice-Hall, 1999.

[44] R. Banner and A. Orda, "Multipath routing algorithms for congestion minimization," *IEEE/ACM Trans. Netw.*, vol. 15, no. 2, pp. 413–424, Apr. 2007.

[45] A. Sprintson, M. Yannuzzi, A. Orda, and X. Masip-Bruin, "Reliable routing with QoS guarantees for multi-domain IP/MPLS networks," in *Proc. IEEE INFOCOM*, 2007, pp. 1820–1828.

[46] S. Secci, J.-L. Rougier, and A. Pattavina, "AS-level source routing for multi-provider connection-oriented services," *Comput. Netw.*, vol. 54, no. 14, pp. 2453–2467, Oct. 2010.

[47] Q. Zhang, M. M. Hasan, X. Wang, P. Palacharla, and M. Sekiya, "Survivable path computation in PCE-based multi-domain networks," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 4, no. 6, pp. 457–467, Jun. 2012.

[48] X. Chen, M. Chamania, A. Jukan, A. C. Drummond, and N. L. S. da Fonseca, "On the benefits of multipath routing for distributed data-intensive applications with high bandwidth requirements and multidomain reach," in *Proc. 7th Annu. CNSR*, 2009, pp. 110–117.

[49] X. Chen, M. Chamania, and A. Jukan, "A backward-compatible interdomain multipath routing framework," in *Proc. IEEE INFOCOM WKSHPS*, 2011, pp. 133–138.

[50] X. Chen, Y. Zhong, and A. Jukan, "Multipath routing in path computation element (PCE): Protocol extensions and implementation," in *Proc. 18th Eur. Conf. OC&i/8th Conf. NOC*, 2013, pp. 75–82.

[51] J. Vasseur and J. L. Roux, "Path computation element (PCE) communication protocol (PCEP)," IETF, Fremont, CA, USA, RFC 5440 (Proposed Standard), Mar. 2009. [Online]. Available: http://www.ietf.org/rfc/rfc5440.txt

[52] D. Specification, "v1. 4," vol. 4, 2003. [Online]. Available: https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-spec-v1.4.0.pdf

[53] H. Newman, A. Barczyk, and M. Bredel, "OLiMPS: Openflow link-layer multipath switching, DOE ASCR NGN PIs Meeting, Sep. 2014. [Online]. Available: http://www.orau.gov/ngnspi2014/presentations/newman_h.pdf

[54] B. Sonkoly, F. Németh, L. Csikor, L. Gulyás, and A. Gulyás, "SDN based testbeds for evaluating and promoting multipath TCP," in *Proc. IEEE ICC*, 2014, pp. 3044–3050.

[55] M. Koerner and O. Kao, "Evaluating SDN based rack-to-rack multi-path switching for data-center networks," *Procedia Comput. Sci.*, vol. 34, pp. 118–125, 2014.

[56] Y. Li and D. Pan, "OpenFlow based load balancing for fat-tree networks with multipath support," in *Proc. 12th IEEE ICC13*, Budapest, Hungary, 2013, pp. 1–5.

[57] A. Tolk *et al.*, "A simulation and emulation study of SDN-based multipath routing for fat-tree data center networks," in *Proc. WSC*, 2014, pp. 3072–3083.

[58] A. Nagata, Y. Tsukiji, and M. Tsuru, "Delivering a file by multipath-multicast on openflow networks," in *Proc. 5th INCoS*, 2013, pp. 835–840.

[59] J. He and J. Rexford, "Toward Internet-wide multipath routing," *IEEE Netw.*, vol. 22, no. 2, pp. 16–21, Mar./Apr. 2008.

[60] Z. Cao, Z. Wang, and E. Zegura, "Performance of hashing-based schemes for Internet load balancing," in *Proc. IEEE INFOCOM*, 2000, vol. 1, pp. 332–341.

[61] C. Hopps, "Analysis of an equal-cost multi-path algorithm," IETF, Fremont, CA, USA, RFC 2992 (Informational), Nov. 2000. [Online]. Available: http://www.ietf.org/rfc/rfc2992.txt

[62] A. Sridharan, R. Guerin, and C. Diot, "Achieving near-optimal traffic engineering solutions for current OSPF/IS–IS networks," *IEEE/ACM Trans. Netw.*, vol. 13, no. 2, pp. 234–247, Apr. 2005.

[63] C. Villamizar, "OSPF optimized multipath (OSPF-OMP)," unpublished.

[64] T. W. Chim, K. L. Yeung, and K.-S. Lui, "Traffic distribution over equal-cost-multi-paths," *Comput. Netw.*, vol. 49, no. 4, pp. 465–475, Nov. 2005.

[65] S. Kandula, D. Katabi, B. Davie, and A. Charny, "Walking the tightrope: Responsive yet stable traffic engineering," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 4, pp. 253–264, Oct. 2005.

[66] S. Kandula, D. Katabi, S. Sinha, and A. Berger, "Dynamic load balancing without packet reordering," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 37, no. 2, pp. 51–62, Apr. 2007.

[67] R. Krishnan and J. A. Silvester, "Choice of allocation granularity in multipath source routing schemes," in *Proc. IEEE INFOCOM*, 1993, pp. 322–329.

[68] V. Paxson, "End-to-end Internet packet dynamics," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 27, no. 4, pp. 139–152, Oct. 1997.

[69] N. M. Piratla and A. P. Jayasumana, "Reordering of packets due to multipath forwarding—An analysis," in *Proc. IEEE ICC*, 2006, vol. 2, pp. 829–834.

[70] K.-C. Leung, V. O. Li, and D. Yang, "An overview of packet reordering in Transmission Control Protocol (TCP): Problems, solutions, and challenges," *IEEE Trans. Parallel Distrib. Syst.*, vol. 18, no. 4, pp. 522–535, Apr. 2007.

[71] J. C. Bennett, C. Partridge, and N. Shectman, "Packet reordering is not pathological network behavior," *IEEE/ACM Trans. Netw.*, vol. 7, no. 6, pp. 789–798, Dec. 1999.

[72] Y. Wang, G. Lu, and X. Li, "A study of Internet packet reordering," in *Information Networking. Networking Technologies for Broadband and Mobile Networks*, H.-K. Kahng and S. Goto, Eds. Berlin, Germany: Springer-Verlag, 2004, pp. 350–359.

[73] X. Zhou and P. Van Mieghem, "Reordering of IP packets in Internet," in *Passive and Active Network Measurement*, C. Barakat and I. Pratt, Eds. Berlin-Germany: Springer-Verlag, 2004, pp. 237–246.

[74] E. Blanton and M. Allman, "On making TCP more robust to packet reordering," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 32, no. 1, pp. 20–30, Jan. 2002.

[75] S. Bohacek, J. P. Hespanha, J. Lee, C. Lim, and K. Obraczka, "TCP-PR: TCP for persistent packet reordering," in *Proc. 23rd Int. Conf. Distrib. Comput. Syst.*, 2003, pp. 222–231.

[76] C. Ma and K.-C. Leung, "Improving TCP reordering robustness in multipath networks," in *Proc. 29th Annu. IEEE Int. Conf. Local Comput. Netw.*, 2004, pp. 409–410.

[77] F. Valera, I. Van Beijnum, A. García-Martínez, and M. Bagnulo, "Multi-path BGP: Motivations and solutions," in *Next-Generation Internet Architectures and Protocols*, B. Ramamurthy, G. N. Rouskas, and K. M. Sivalingam, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[78] S. M. Bellovin, "Security problems in the TCP/IP protocol suite," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 19, no. 2, pp. 32–48, Apr. 1989.

[79] R. Albrightson, J. Garcia-Luna-Aceves, and J. Boyle, "EIGRP—A fast routing protocol based on distance vectors," in *Proc. Networld/Interop*, 1994, vol. 94, pp. 136–147.

[80] T. Przygienda, "M-ISIS: Multi Topology (MT) routing in Intermediate System to Intermediate Systems (IS–ISs)," IETF, Fremont, CA, USA, RFC 5120, Feb. 2008.

[81] M. Menth and R. Martin, "Network resilience through multi-topology routing," in *Proc. 5th Int. Workshop Des. Reliable Commun. Netw.*, 2005, pp. 271–277.

[82] A. Kvalbein and O. Lysne, "How can multi-topology routing be used for intradomain traffic engineering?" in *Proc. SIGCOMM Workshop Internet Netw. Manage.*, 2007, pp. 280–284.

[83] Y. Rekhter and T. Li, "A border gateway protocol 4 (BGP-4)," IETF, Fremont, CA, USA, RFC 1771, Mar. 1995.

[84] J. Scudder, A. Retana, D. Walton, and E. Chen, "Advertisement of Multiple Paths in BGP," 2013.

[85] W. Xu and J. Rexford, "MIRO: Multi-path interdomain routing," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 4, pp. 171–182, Oct. 2006.

[86] I. Van Beijnum, J. Crowcroft, F. Valera, and M. Bagnulo, "Loop-freeness in multipath BGP through propagating the longest path," in *Proc. IEEE ICC Workshops*, 2009, pp. 1–6.

[87] S. Agarwal, A. Nucci, and S. Bhattacharyya, "Controlling hot potatoes in intradomain traffic engineering," SPRINT ATL, Burlingame, CA, USA, Res. Rep. RR04-ATL-070677, Jul. 2004.

[88] Cisco, San Jose, CA, USA, "BGP best path selection algorithm." [Online]. Available: http://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/13753-25.pdf

[89] Juniper, Laredo, TX, USA, "Understanding BGP multipath." [Online]. Available: http://www.juniper.net/documentation/en_US/junos14.1/topics/concept/bgp-multipath-understanding.html

[90] M. Motiwala, M. Elmore, N. Feamster, and S. Vempala, "Path splicing," in *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, 2008, pp. 27–38.

[91] D. Zhu, M. Gritter, and D. R. Cheriton, "Feedback based routing," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 33, no. 1, pp. 71–76, 2003.

[92] X. Yang, D. Clark, and A. W. Berger, "NIRA: A new inter-domain routing architecture," *IEEE/ACM Trans. Netw.*, vol. 15, no. 4, pp. 775–788, Aug. 2007.

[93] P. Godfrey, I. Ganichev, S. Shenker, and I. Stoica, "Pathlet routing," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 111–122, 2009.

[94] H. T. Kaur *et al.*, "BANANAS: An evolutionary framework for explicit and multipath routing in the Internet," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 33, no. 4, pp. 277–288, 2003.

[95] D. Andersen, H. Balakrishnan, F. Kaashoek, and R. Morris, "Resilient overlay networks," *ACM SIGOPS Oper. Syst. Rev.*, vol. 35, no. 5, pp. 131–145, Dec. 2001.

[96] J. M. Camacho, A. García-Martínez, M. Bagnulo, and F. Valera, "BGP-XM: BGP eXtended multipath for transit autonomous systems," *Comput. Netw.*, vol. 57, no. 4, pp. 954–975, Mar. 2013.

[97] E. Altman, T. Boulogne, R. El-Azouzi, T. Jiménez, and L. Wynter, "A survey on networking games in telecommunications," *Comput. Oper. Res.*, vol. 33, no. 2, pp. 286–311, Feb. 2006.

[98] S. Secci, J. Rougier, A. Pattavina, F. Patrone, and G. Maier, "Peering equilibrium multipath routing: A game theory framework for Internet peering settlements," *IEEE/ACM Trans. Netw.*, vol. 19, no. 2, pp. 419–432, Apr. 2011.

[99] S. Secci, K. Liu, and B. Jabbari, "Efficient inter-domain traffic engineering with transit-edge hierarchical routing," *Comput. Netw.*, vol. 57, no. 4, pp. 976–989, Mar. 2013.

[100] S. Mao, D. Bushmitch, S. Narayanan, and S. S. Panwar, "MRTP: A multiflow real-time transport protocol for ad hoc networks," *IEEE Trans. Multimedia*, vol. 8, no. 2, pp. 356–369, Apr. 2006.

[101] V. Singh, T. Karkkainen, J. Ott, S. Ahsan, and L. Eggert, "Multipath RTP (MPRTP)," IETF, Fremont, CA, USA, Internet-Draft, Tech. Rep., Jul. 2012.

[102] V. Jacobson, R. Frederick, S. Casner, and H. Schulzrinne, "RTP: A transport protocol for real-time applications," IETF, Fremont, CA, USA, RFC 3550, Jul. 2003.

[103] V. Singh, S. Ahsan, and J. Ott, "MPRTP: Multipath considerations for real-time media," in *Proc. 4th ACM Multim. Syst. Conf.*, 2013, pp. 190–201.

[104] C. Perkins and V. Singh, "Multimedia congestion control: Circuit breakers for unicast RTP sessions," 2015. [Online]. Available: https://tools.ietf.org/html/draft-ietf-avtcore-rtp-circuit-breakers-10

[105] B. Cohen, "The BitTorrent protocol specification," 2008. [Online]. Available: http://www.bittorrent.org/beps/bep_0003.html

[106] M. Day, B. Cain, G. Tomlinson, and P. Rzewski, "A model for content internetworking (CDI)," IETF, Fremont, CA, USA, RFC 3466, Feb. 2003.

[107] J. Apostolopoulos, T. Wong, W.-T. Tan, and S. Wee, "On multiple description streaming with content delivery networks," in *Proc. IEEE INFOCOM*, 2002, vol. 3, pp. 1736–1745.

[108] H. Sivakumar, S. Bailey, and R. L. Grossman, "PSockets: The case for application-level network striping for data intensive applications using high speed wide area networks," in *Proc. ACM/IEEE Conf. Supercomput.*, 2000, p. 38.

[109] T. J. Hacker, B. D. Athey, and B. Noble, "The end-to-end performance effects of parallel TCP sockets on a lossy wide-area network," in *Proc. IPDPS, Abstracts and CD-ROM*, 2001, pp. 1–10.

[110] S. Gouache, G. Bichot, and C. Howson, "Distributed & adaptive HTTP streaming," in *Proc. IEEE ICME*, 2011, pp. 1–6.

[111] J. Kim *et al.*, "Multi-source multipath HTTP (mHTTP): A proposal," in *Proc. ACM Int. Conf. Meas. Model. Comput. Syst.*, 2014, pp. 583–584.

[112] N. F. Maxemchuk, "Dispersity routing," in *Proc. ICC*, 1975, vol. 75, pp. 41–10.

[113] N. F. Maxemchuk, "Dispersity routing in high-speed networks," *Comput. Netw. ISDN Syst.*, vol. 25, no. 6, pp. 645–661, Jan. 1993.

[114] "Multipath TCP resources." [Online]. Available: http://datatracker.ietf.org/wg/mptcp/documents/

[115] C. Huitema, "Multi-homed TCP draft-huitema-multi-homed-0," IETF, Fremont, CA, USA, May 1995.

[116] H.-Y. Hsieh and R. Sivakumar, "pTCP: An end-to-end transport layer protocol for striped connections," in *Proc. 10th IEEE Int. Conf. Netw. Protocols*, 2002, pp. 24–33.

[117] K. Rojviboonchai and A. Hitoshi, "An evaluation of multi-path transmission control protocol (M/TCP) with robust acknowledgement schemes," *IEICE Trans. Commun.*, vol. 87, no. 9, pp. 2699–2707, 2004.

[118] M. Zhang, J. Lai, A. Krishnamurthy, L. Peterson, and R. Wang, "A transport layer approach for improving end-to-end performance and robustness using redundant paths," in *Proc. USENIX Annu. Tech. Conf., General Track*, 2004, pp. 99–112.

[119] Y. Dong, N. Pissinou, and J. Wang, "Concurrency handling in TCP," in *Proc. 5th Annu. Conf. Commun. Netw. Services Res.*, 2007, pp. 255–262.

[120] Y. Hasegawa *et al.*, "Improved data distribution for multipath TCP communication," in *Proc. IEEE GLOBECOM*, 2005, vol. 1, p. 271.

[121] C. Raiciu, M. Handley, and D. Wischik, "Coupled congestion control for multipath transport protocols," IETF, Fremont, CA, USA, RFC 6356, Oct. 2011.

[122] C. Raiciu *et al.*, "Improving datacenter performance and robustness with multipath TCP," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 266–277, Aug. 2011.

[123] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 63–74, 2008.

[124] C. Villamizar and T. Li, "IS–IS optimized multipath (ISIS-OMP)," DRAFT-IETF-ISIS-OMP-02, Internet Draft, 1998.

[125] A. Khanna and J. Zinky, "The revised ARPANET routing metric," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 19, no. 4, pp. 45–56, 1989.

[126] X. Yang and D. Wetherall, "Source selectable path diversity via routing deflections," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 4, pp. 159–170, Oct. 2006.

[127] D. Farinacci, V. Fuller, D. Meyer, and D. Lewis, "The locator/ID separation protocol," IETF, Fremont, CA, USA, RFC 6830, Jan. 2013.

[128] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, "The nature of data center traffic: Measurements & analysis," in *Proc. 9th ACM SIGCOMM Conf. Internet Meas. Conf.*, 2009, pp. 202–208.

[129] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in *Proc. 10th ACM SIGCOMM Conf. Internet Meas.*, 2010, pp. 267–280.

[130] F. P. Tso and D. P. Pezaros, "Improving data center network utilization using near-optimal traffic engineering," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 6, pp. 1139–1148, Jun. 2013.

[131] T. Benson, A. Anand, A. Akella, and M. Zhang, "MicroTE: Fine grained traffic engineering for data centers," in *Proc. 7th Conf. Emerg. Netw. Exp. Technol.*, 2011, pp. 1–12.

[132] D. Xu, M. Chiang, and J. Rexford, "Link-state routing with hop-by-hop forwarding can achieve optimal traffic engineering," *IEEE/ACM Trans. Netw.*, vol. 19, no. 6, pp. 1717–1730, Dec. 2011.

[133] *IEEE Standards for Local and Metropolitan Area Networks—Virtual Bridged Local Area Networks—Amendment 3: Multiple Spanning Trees*, IEEE Std. 802.1s-2002 (Amendment to IEEE Std 802.1Q, 1998 Edition), 2002.

[134] S. Sharma, K. Gopalan, S. Nanda, and T.-C. Chiueh, "Viking: A multi-spanning-tree Ethernet architecture for metropolitan area and cluster networks," in *Proc. IEEE INFOCOM*, 2004, vol. 4, pp. 2283–2294.

[135] *IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks*, IEEE Std. 802.1Q-1998, 1999.

[136] J. Mudigonda, P. Yalagandula, M. Al-Fares, and J. C. Mogul, "SPAIN: COTS data-center Ethernet for multipathing over arbitrary topologies," in *Proc. NSDI*, 2010, pp. 265–280.

[137] D. Fedyk, P. Ashwood-Smith, D. Allan, A. Bragg, and P. Unbehagen, "IS–IS extensions supporting IEEE 802.1aq shortest path bridging," IETF, Fremont, CA, USA, RFC 6329, Apr. 2012.

[138] D. Allan *et al.*, "Shortest path bridging: Efficient control of larger Ethernet networks," *IEEE Commun. Mag.*, vol. 48, no. 10, pp. 128–135, Oct. 2010.

[139] J. Touch and R. Perlman, "Transparent interconnection of lots of links (Trill): Problem and applicability statement," IETF, Fremont, CA, USA, RFC 5556, May 2009.

[140] A. Greenberg *et al.*, "VL2: A scalable and flexible data center network," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, 2009, pp. 51–62.

[141] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, "Hedera: Dynamic flow scheduling for data center networks," in *Proc. NSDI*, 2010, vol. 10, pp. 1–15.

[142] A. R. Curtis *et al.*, "Devoflow: Scaling flow management for high-performance networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 254–265, Aug. 2011.

[143] A. Elwalid, C. Jin, S. Low, and I. Widjaja, "MATE: MPLS adaptive traffic engineering," in *Proc. IEEE INFOCOM*, 2001, vol. 3, pp. 1300–1309.

[144] Y. Seok, Y. Lee, Y. Choi, and C. Kim, "Dynamic constrained multipath routing for MPLS networks," in *Proc. IEEE Comput. Commun. Netw.*, 2001, pp. 348–353.

[145] Y. Lee, Y. Seok, Y. Choi, and C. Kim, "A constrained multipath traffic engineering scheme for MPLS networks," in *Proc. IEEE ICC*, 2002, vol. 4, pp. 2431–2436.

[146] H. Saito, Y. Miyao, and M. Yoshida, "Traffic engineering using multiple multipoint-to-point LSPs," in *Proc. IEEE INFOCOM*, 2000, vol. 2, pp. 894–901.

[147] Y. Wang and Z. Wang, "Explicit routing algorithms for Internet traffic engineering," in *Proc. 8th Int. Conf. Comput. Commun. Netw.*, 1999, pp. 582–588.

[148] M. Menth, A. Reifert, and J. Milbrandt, "Self-protecting multipaths—A simple and resource-efficient protection switching mechanism for MPLS networks," in *Networking*, N. Mitrou, K. Kontovasilis, G. N. Rouskas, I. Iliadis, and L. Merakos, Eds. Berlin, Germany: Springer-Verlag, 2004, pp. 526–537.

[149] H. Suzuki and F. A. Tobagi, "Fast bandwidth reservation scheme with multi-link and multi-path routing in ATM networks," in *Proc. IEEE INFOCOM*, 1992, pp. 2233–2240.

[150] T. T. Lee and S. C. Liew, "Parallel communications for ATM network control and management," in *Proc. IEEE GLOBECOM*, 1993, pp. 442–446.

[151] J. Sole-Pareta, D. Sarkar, J. Liebeherr, and I. F. Akyildiz, "Adaptive multipath routing of connectionless traffic in an ATM network," in *Proc. IEEE ICC*, 1995, vol. 3, pp. 1626–1630.

[152] R. Niranjan Mysore *et al.*, "Portland: A scalable fault-tolerant layer 2 data center network fabric," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 39–50, Oct. 2009.

[153] X. Chen, A. Jukan, A. C. Drummond, and N. L. Da Fonseca, "A multipath routing mechanism in optical networks with extremely high bandwidth requests," in *Proc. IEEE GLOBECOM*, 2009, pp. 1–6.

[154] L. Ruan and N. Xiao, "Survivable multipath routing and spectrum allocation in OFDM-based flexible optical networks," *J. Opt. Commun. Netw.*, vol. 5, no. 3, pp. 172–182, Mar. 2013.

[155] X. Chen, Y. Zhong, and A. Jukan, "Multipath routing in elastic optical networks with distance-adaptive modulation formats," in *Proc. IEEE ICC*, 2013, pp. 3915–3920.

[156] X. Chen, A. Jukan, and A. Gumaste, "Multipath de-fragmentation: Achieving better spectral efficiency in elastic optical path networks," in *Proc. IEEE INFOCOM*, 2013, pp. 390–394.

[157] "Network node interface for the synchronous digital hierarchy (SDH)," ITU, Geneva, Switzerland, ITU-T Recommendation G.707, Dec. 2003.

[158] R. Ramaswami, K. Sivarajan, and G. Sasaki, *Optical Networks: A Practical Perspective*. San Mateo, CA, USA: Morgan Kaufmann, 2009.

[159] E. Hernandez-Valencia, M. Scholten, and Z. Zhu, "The generic framing procedure (GFP): An overview," *IEEE Commun. Mag.*, vol. 40, no. 5, pp. 63–71, May 2002.

[160] "Link capacity adjustment scheme (LCAS) for virtual concatenated signals," ITU, Geneva, Switzerland, ITU-T Recommendation G.7042, Feb. 2004.

[161] C. Ou, L. H. Sahasrabuddhe, K. Zhu, C. U. Martel, and B. Mukherjee, "Survivable virtual concatenation for data over SONET/SDH in optical transport networks," *IEEE/ACM Trans. Netw.*, vol. 14, no. 1, pp. 218–231, Feb. 2006.

[162] S. Acharya, B. Gupta, P. Risbood, and A. Srivastava, "PESO: Low overhead protection for Ethernet over SONET transport," in *Proc. IEEE INFOCOM*, 2004, vol. 1, pp. 1–11.

[163] S. S. Ahuja, T. Korkmaz, and M. Krunz, "Minimizing the differential delay for virtually concatenated Ethernet over SONET systems," in *Proc. 13th ICCCN*, 2004, pp. 205–210.

[164] M. Jinno *et al.*, "Multiflow optical transponder for efficient multilayer optical networking," *IEEE Commun. Mag.*, vol. 50, no. 5, pp. 56–65, May 2012.

[165] J. Y. Yen, "Finding the K shortest loopless paths in a network," *Manag. Sci.*, vol. 17, no. 11, pp. 712–716, 1971.

[166] J. Song, S. Kim, M. Lee, H. Lee, and T. Suda, "Adaptive load distribution over multipath in NEPLS networks," in *Proc. IEEE ICC*, 2003, vol. 1, pp. 233–237.

[167] W. T. Zaumen and J. Garcia-Luna-Aceves, "Loop-free multipath routing using generalized diffusing computations," in *Proc. IEEE INFOCOM*, 1998, vol. 3, pp. 1408–1417.

[168] K. Lee, A. Toguyeni, A. Noce, and A. Rahmani, "Comparison of multipath algorithms for load balancing in a MPLS network," in *Information Networking. Convergence in Broadband and Mobile Networking*. Berlin, Germany: Springer-Verlag, 2005, pp. 463–470.

[169] M. Coudron, S. Secci, G. Pujolle, P. Raad, and P. Gallard, "Cross-layer cooperation to boost multipath TCP performance in cloud networks," in *Proc. IEEE 2nd Int. Conf. CloudNet*, 2013, pp. 58–66.

[170] A. S. Anghel, R. Birke, D. Crisan, and M. Gusat, "Cross-layer flow and congestion control for datacenter networks," in *Proc. 3rd Workshop Data Center-Converged Virtual Ethernet Switching*, 2011, pp. 44–62.

[171] M. Coudron *et al.*, "Boosting cloud communications through a cross-layer multipath protocol architecture," in *Proc. IEEE SDN4FNS*, 2013, pp. 1–8.

[172] M. Woo, "High-energy physicists smash records for network data transfer," Caltech Media Relations, Pasadena, CA, USA, Nov. 2012. [Online]. Available: http://www.caltech.edu/news/high-energy-physicists-smash-records-network-data-transfer-37565

[173] P. Lappas, "SDN use case: Multipath TCP at Caltech and CERN," Project Floodlight, Dec. 2012. [Online]. Available: http://www.projectfloodlight.org/blog/2012/12/03/sdn-use-case-multipath-tcp-at-caltech-and-cern/

[174] R. Van Der Pol *et al.*, "Multipathing with MPTCP and OpenFlow," in *Proc. SCC*, 2012, pp. 1617–1624.

[175] X. Chen, A. Jukan, and A. Gumaste, "Optimized parallel transmission in elastic optical networks to support high-speed ethernet," *J. Lightw. Technol.*, vol. 32, no. 2, pp. 228–238, Jan. 2014.

[176] X. Chen, A. Jukan, and M. Médard, "Linear network coding and parallel transmission increase fault tolerance and optical reach," in *Proc. IEEE ICC*, 2015, pp. 1–6.

[177] H. Han, S. Shakkottai, C. V. Hollot, R. Srikant, and D. Towsley, "Multipath TCP: A joint congestion control and routing scheme to exploit path diversity in the Internet," *IEEE/ACM Trans. Netw.*, vol. 14, no. 6, pp. 1260–1271, Dec. 2006.

[178] C. Raiciu, D. Wischik, and M. Handley, "Practical congestion control for multipath transport protocols," Univ. College London, London, U.K., Tech. Rep., 2009.

[179] F. P. Kelly, A. K. Maulloo, and D. K. Tan, "Rate control for communication networks: Shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, pp. 237–252, 1998.

[180] F. Kelly and T. Voice, "Stability of end-to-end algorithms for joint routing and rate control," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 2, pp. 5–12, Apr. 2005.

[181] J. He, M. Bresler, M. Chiang, and J. Rexford, "Towards robust multi-layer traffic engineering: Optimization of congestion control and routing," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 5, pp. 868–880, Jun. 2007.

[182] A. F. T. Committee, "Private network-network interface specification Version 1.0 (PNNI 1.0)," ATM Forum, Tech. Rep. af-pnni-0055.000, Mar. 1996.

[183] I. Cidon, R. Rom, and Y. Shavitt, "Multi-path routing combined with resource reservation," in *Proc. IEEE INFOCOM*, 1997, vol. 1, pp. 92–100.

[184] I. Cidon, R. Rom, and Y. Shavitt, "Analysis of multi-path routing," *IEEE/ACM Trans. Netw.*, vol. 7, no. 6, pp. 885–896, Dec. 1999.

[185] N. M. Piratla, A. P. Jayasumana, and A. A. Bare, "Reorder density (RD): A formal, comprehensive metric for packet reordering," in *NETWORKING 2005. Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems*, R. Boutaba, K. C. Almeroth, R. Puigjaner, S. X. Shen, and J. P. Black, Eds. Berlin, Germany: Springer-Verlag, 2005, pp. 78–89.

[186] N. M. Piratla, A. P. Jayasumana, and A. Bare, "A comparative analysis of packet reordering metrics," in *Proc. IEEE/ACM 1st Int. Conf. COMSWARE, New Delhi, India*, 2006, pp. 1–10.

[187] K.-C. Leung and V. O. Li, "Flow assignment and packet scheduling for multipath routing," *J. Commun. Netw.*, vol. 5, no. 3, pp. 230–239, Sep. 2003.

[188] N. Gogate and S. S. Panwar, "On a resequencing model for high speed networks," in *Proc. IEEE INFOCOM*, 1994, pp. 40–47.

[189] K.-C. Leung and V. Li, "A resequencing model for high speed networks," in *Proc. IEEE ICC*, 1999, vol. 2, pp. 1239–1243.

[190] Z. Wang and J. Crowcroft, "Analysis of shortest-path routing algorithms in a dynamic network environment," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 22, no. 2, pp. 63–71, Apr. 1992.

[191] K. Varadhan, R. Govindan, and D. Estrin, "Persistent route oscillations in inter-domain routing," *Comput. Netw.*, vol. 32, no. 1, pp. 1–16, 2000.

[192] X. Liu, S. Mohanraj, M. Pioro, and D. Medhi, "Multipath routing from a traffic engineering perspective: How beneficial is it?" in *Proc. IEEE 22nd ICNP*, 2014, pp. 143–154.

[193] A. Al-Dhaher, T. Anjali, and A. Fortin, "Achieving distributed buffering in multi-path routing using fair allocation," in *Proc. IEEE Int. Conf. EIT*, 2011, pp. 1–6.

[194] J. Santos, J. Pedro, P. Monteiro, and J. Pires, "Optimized routing and buffer design for optical transport networks based on virtual concatenation," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 3, no. 9, pp. 725–738, Sep. 2011.

[195] X. Chen, A. Engelmann, A. Jukan, and M. Medard, "Linear network coding reduces buffering in high-speed Ethernet parallel transmission systems," *IEEE Commun. Lett.*, vol. 18, no. 4, pp. 636–639, Apr. 2014.

[196] M. Li, A. Lukyanenko, and Y. Cui, "Network coding based multipath TCP," in *Proc. IEEE INFOCOM WKSHPS*, 2012, pp. 25–30.

Sandeep Kumar Singh (S'15) received the M.S. degree in electrical engineering from the Indian Institute of Technology Madras, Chennai, India. He is currently working toward the Ph.D. degree in communication networks with the Technische Universität Carolo-Wilhelmina zu Braunschweig, Braunschweig, Germany. His research interests include optical networks, traffic engineering in data centers, and stochastic analysis.

Tamal Das received the M.Tech. degree from the Indian Institute of Technology (IIT) Delhi, New Delhi, India, and the Ph.D. degree from the IIT Bombay, Mumbai, India. He is currently a Postdoctoral Researcher at the Institut für Datentechnik und Kommunikationsnetze, Technische Universität Carolo-Wilhelmina zu Braunschweig, Braunschweig, Germany. He has authored over 20 high-quality scientific publications. His research interests are in stochastic analysis, telecommunication networks and network algorithms. Dr. Das is the Project Manager of FP7 EU-Project PACE (ict-pace.net). He serves on the TPC of Conference of Telecommunication, Media and Internet Techno-Economics (CTTE) 2015. He was a recipient of the IEEE ANTS 2010 Best Paper Award.

Admela Jukan received the Dipl.-Ing. degree from the Fakultet Elektrotehnike i Racunarstva (FER), Zagreb, Croatia, the M.Sc. degree in information technologies from the Politecnico di Milano, Milan, Italy, and the Dr.Tech. degree (*cum laude*) in electrical and computer engineering from the Technische Universität Wien, Wien, Austria. She is a Chair Professor of Communication Networks at the Technische Universität Carolo-Wilhelmina zu Braunschweig (TU Braunschweig), Braunschweig, Germany. She is also a Departmental Coordinator of the International Student Exchange ERASMUS, TU Braunschweig. Prof. Jukan is an Elected Distinguished Lecturer of the IEEE Communications Society (2015–2017). She has chaired and co-chaired several international conferences, including IFIP ONDM, IEEE ANTS, IEEE ICC, and IEEE GLOBECOM. She is an Elected Chair of the IEEE Optical Network Technical Committee, ONTC (2014–2015). She serves as an Associate Technical Editor of the IEEE COMMUNICATIONS MAGAZINE and as a Senior Editor of the IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS. She is a Co-Editor-in-Chief of the *Journal on Optical Switching and Networking* (Elsevier). She is the Coordinating Principal Investigator of the FP7 EU-Project PACE (ict-pace.net), focusing on innovation in next-generation network management systems. She was a recipient of an Award of Excellence for the BMBF/CELTIC project 100 Gb Ethernet and the IBM Innovation Award (2009).