INTAREA Working Group S. Bryant
Internet-Draft University of Surrey 5/6GIC
Intended status: Informational U. Chunduri
Expires: July 28, 2022 Intel
T. Eckert
A. Clemm
Futurewei Technologies Inc.
January 24, 2022

Forwarding Layer Problem Statement
draft-bryant-arch-fwd-layer-ps-04

Abstract

This document considers the problems that need to addressed in IP in order to address the use cases and new network services described in draft-bryant-arch-fwd-layer-uc-00.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

There is an emerging set of new requirements that exceed the network and transport services of the current Internet, which only delivers "best effort" service. While many controlled or private networks include further services, such as other DiffServ QoS in addition to best effort and traffic engineering with bandwidth guarantees, the solutions used today only support walled gardens and are thus not available to application service providers and consumers across the Internet.

The uses cases and service needs that are foreseen as necessary for deployment in the medium future are described in [I-D.bryant-arch-fwd-layer-uc].

The purpose of this document is to examine the shortcomings that the existing network and transport layer protocols a well as their associated control plane need to overcome to meet these needs.

The IETF is the body responsible for the long term evolution of the IP protocol suit, but is missing a work track to discuss the long-term Internet network architecture evolution. In particular it lacks a programme for the long term evolution of IP itself.

Approximately 30 years ago, the IETF started a process to revolutionize the IPv4 [RFC0791] Internet Protocol. In this process, researchers, industry, and service providers got together, and brought up a number of new proposals, and worked toward a successor to IPv4, which became IPv6 [RFC2460] and later [RFC8200].

30 years later, there is heavy resistance to anything more than minor incremental evolutions to IPv6. There are a number of reasons for this ranging from opinions that all future IP needs can be met through minor incremental evolutions to fears that major proposals for innovation at the IP would be an unwelcome disrupter to the current business of the vendors or the service providers.

The authors take no position on the scale of the problem or the difficulties of deploying any solutions at scale in the Internet. What we seek to do is to establish the scope and nature of the problem. A decision on which aspects of the problem are economically tractable is out of scope of this text, but technologies to support monetization are not.

As a problem statement, this documents goal is to not propose or promote specific solutions to the problems raised. Instead it uses references to not Internet adopted, but proposed or existing solutions only as example evidence that the described problem can actually be solved.

Because the document does not propose specific solutions, it also does not attempt to structure the problem description in a way that would identify sub-set of problems to be resolved by specific solution components.

The purpose of this text is thus to stimulate discussion on the emerging needs of the forwarding layer and to start the process of determining how they are best satisfied within the IETF protocol suite.

2. Forwarding Layer

The term "forwarding layer" is used in this work because none of the standard terms encompass the parts of the network stack that need attention to address the needs of the applications that are foreseen.

It is possible that development work will need to reach down to layer 2.5 in order to ensure that packets are handled correctly down to the physical layer. The MAC layer is quite sophisticated and includes its own switching function so we need to be sure that the good work done in the network layer is not undone lower down the stack. Equally it is possible that development work will need to reach into the transport layer to address new approaches to congestion, and to ensure that the network layer understands the requirements placed on it by the application. An open mind is needed on the boundaries of the layers as they exist today when analyzing the consequential network changes needed to support the evolving application space.

In the network layer itself, this document is only concerned with the forwarding component, not path selection or the other components of routing.

Thus, we use the term forwarding layer to describe the scope of the stack that this document addresses.

3. Underlying New Requirements

3.1. Better than Best Effort

The current Internet is essentially of best-effort system, but future applications require high-precision KPIs on throughput, latency and packet loss for industrial manufacturing, control, automation, and machine-to-machine communications.

The emerging use cases for networks require deployment of capabilities that are beyond best effort. Best effort networks can do remarkably well by simply throwing bandwidth at the problem and lightly loading the network. For the case where a greater capability is needed the IETF has invested effort in deterministic networking (DN) [RFC8655]. Whilst DN is an improvement over best effort it is still fundamentally a best effort service with enhancement to improved the probability of a packet not being delayed or lost due to congestion. It is an after the fact enhancement to the method of operation of what is a largely unmodified data plane. I the case of MPLS [RFC8964] there is some assistance from the PREOF function, but IP runs the standard data plane and relies entirely on special case packet selection queue management. It is thus an after-the-fact

enhancement to a minimally changed data plane restricted to a single network domain.

With upcoming Cellular technologies (5G/B5G) there is a need for Service Providers to expand the type of customers for metropolitan size networks to address their better than best-effort traffic needs.

DetNet has been proposed to support this, however:

- o Only some aspects of DetNet currently only run on top of current IP/IPv6.

- o DetNet service is constrained: It only supports constant bit rate (CBR), reserved bandwidth. It does not support flexible bandwidth. The notion of contracts in a future development of the forwarding layer will support more flexible managed bandwidth and managed latency contracts for traffic.

3.2. Efficient Packet Design

The ratio of useful data in the payload to overhead has a direct financial impact on communication links; these links are of finite capacity and hence have a finite cost-per-unit-data that can be calculated. The capacity used to transport information as compared to the overhead which is unavailable for use by a customer, but required to transmit is often expresses as a good-put efficiency and can be related to cost to transmit payload data.

- o There is a need to support large number of low power user equipment (UE) devices (low-power IoTs) connecting through various radio networks (LTE/5G/B5G) where spectral efficiency is needed. This needs to be achieved without header compression techniques like as [RFC6282] since, compression can result in additional processing and energy consumption overhead.

- o The handling network protocol headers, requires that portions of each packet be held in memory or buffer structures; the more levels of information which need to be held for processing by network nodes, the more memory space will be required, and this directly effects the cost of operation and cost of manufacture/ provision of such equipment.

On the other hand, in various non-constrained environments where various network layer functionalities are desired, there are different set of requirements. For example:

- o Segment Routing over IPv6 (SRv6) parameter encoding [RFC8986] in the SRv6 SID [RFC8754] is limited by the prefix portion of the IPv6 address.

- o In Identifier Locator Addressing (ILA), the identifier (ID) portion of the address length is limited because of 128 bits limit.

3.3. Forwarding Identifiers

Developments in IPv6 {{RFC8986} formalize a trend that has been happening for a long time: the morphing of network layer addresses into forwarding identifiers (FI). However, constraining FIs to a fixed size ill serves the development of the forwarding layer. There are clear cases as illustrated above where it would be useful to have shorter network layer addresses. Equally we can see that there will be future cases where 128 bits may be insufficient to specify a forwarding operation. The requirement is thus to formally introduce the concept of forwarding identifiers in place of network layer addresses, and use a forwarding identifier construct that supports multiple semantics and multiple, possibly fully variable, lengths.

There is further discussion on this point in Section 4.1.2.

3.4. Operational visibility

Network operators require facilities that let them better understand and fine tune detailed network behavior. These features are hard to retrofit with current IP/IPv6.

The rise of machine learning has led to the expectation of being able to better optimize networks This in turn leads to the increase of network telemetry as a source of data to base these systems on. In-Situ OAM (IOAM) [I-D.ietf-ippm-ioam-data] represents one of the latest developments in that space, allowing the data plane to piggy-back telemetry data onto individual packets in order to diagnose and fine-tune service levels such as latency or jitter. However, there are several issues with this approach:

- o MTU issues limit amount of data that can be obtained. With IOAM packet size increases with number of data items and number of hops.

- o The data that can be obtained is very limited.

- o The OAM data volume can easily exceeds that of production traffic which is wasteful

- There is no ability to aggregate OAM data, or make context dependent OAM collection.

- Integration with other solutions such as DetNet is unclear.

While useful, IOAM exposes the limits of what add-on solutions can provide. Solutions that provide visibility at the level of flows or that provide automatic verification of Service Level Objectives are missing entirely.

3.5. Holistic Solution

It needs to be recognized that it will not be sufficient for solutions to support new services and capabilities one at a time and independently from one another. For example, better-than-best-effort, operational visibility, and efficient packet design should go together, without leading to additional integration problems ore requiring users to make a choice.

A piecemeal approach, in which solutions for any one particular problem are developed and emerge one at a time, results in a fragmented solution which gets progressively more difficult to integrate with components previously designed. Thus it is better if solutions are holistic and be able to support new services and capabilities in integrated fashion and simultaneously with each other.

We therefore need to identify an elegant approach that is simple and naturally extensible to address problems that we do not yet conceive as requiring addressing.

Any such solution needs to be intrinsically secure and yet be able to support security without privacy and privacy without security.

4. Existing Protocol, Layering Challenges and Gaps

Despite IPv4 still having a large user base, and having a number of useful properties the IETF has abandoned future development of IPv4 as a way to force the deployment of IPv6. For example, in terms of traffic steering the segment routing could have usefully been applied to IPv4 to support network operators that wished to retain IPv4 as their preferred internal protocol.

Given the gaps in each of the existing network layer protocols the IETF may wish to look at the design of a protocol that both fills the gaps and unifies its three existing network layer protocols: IPv4, IPv6 and MPLS.

Additionally there is a clear need for a more sophisticated approach to indicating the required quality of service that a packet, or flow, needs in an IP network.

4.1. Challenges with IPv6

4.1.1. The End-to-End Model

IPv6 and specifically [RFC8200] was designed to fit within an Internet architecture centered around the end-to-end model with "Internet Paths" potentially passing through one or more networks without any relationship to the endpoints of a communication such as most so-called transit-AS. As history already from IPv4 had shown, anything more than the most simple per-hop processing options can cause interoperability issues. In result, [RFC8200] has drastically limited such per-hop processing options.

Two core restrictions of RFC8200 are the following:

- o Restrictions on extension headers (EH): EHs must never be deleted or changed in size by any node on the path the packet takes. Intermediate nodes are only expected to examine these headers (if they are configured to do so). Implementations cannot expect intermediate nodes to examine, or act on, except for hop-by-hop header (section 4.8 of [RFC8200]).

At the time of writing this is an area of considerable active discussion in the IETF 6MAN and SPRING working groups. The issues that arise from allowing unrestricted insertion, deletion or modification of EHs are for example:

- o Breakage of path MTU discovery
- o Impact on the Authentication Header protocol
- o Inability to return ICMP error messages to the correct node.

See Section 4.1.1.1 for further discussion.

- o No new hop-by-hop headers (HBH) in IPV6: No new EHs that require hop-by-hop behavior should be defined (section 4 of [RFC8200]) - the only EH that has hop-by-hop behavior is the Hop-by-Hop Options header. The only alternative available to the designer is instead to use destination headers (section 6.8 of [RFC8200]).

4.1.1.1. IPv6 For Controlled Networks

While [RFC8200] is a conservative set of requirements to enable proliferation of the target use case of "Internet Paths", the same set of requirements limit the flexibility of IPv6 unnecessarily when it is used in controlled networks where the constraints and interoperability issues for "Internet Paths" do not equally apply, for example the deployment scenarios described in Sections "Embedded Service" and "Embedded Global Service" of [I-D.bryant-arch-fwd-layer-uc].

One typical type of controlled networks are service providers (SP) where SRv6 is used as the architecture within the SP network.

- o IPv6 extension headers can not be added on a midpoint. Any addition/change requires an encapsulation where another IPv6 header with optional SRH extension header is prepended to the carried IPv6 packet. This is expensive in terms of packet MTU, and in terms of packet buffer requirements at the ends of the provider path which can be an economic issue in cost sensitive network segments.

- o The requirement to encapsulate instead of being allowed to add an EH along the path stems from the desire to isolate any header changes from Path MTU Discovery (PMTUD). This is a necessary complexity when traversing uncontrolled hops across the Internet, but it is unnecessary overhead when only passing through controlled hops. In MPLS and SR-MPLS, the MPLS header size is not included in the MTU available to the MPLS payload, instead the network is managed such that the maximum MPLS header size plus the available payload MTU is always smaller that the encapsulating L2 frame MTU. In IPv6 instead, the encapsulating and decapsulating would logically have to perform signaling for PMTUD (unnecessarily).

- o Because of the authorization header (AH) [RFC4302] and OAM concerns, [RFC8200] likewise prohibits removing extension headers or fields thereof on hops along the path, requiring for example more complex packet parsers. In SR-MPLS it is possible to simply remove the top SID on a node that has processed it, in SRv6 it is instead necessary to look up an offset field in the SRH and, read the appropriate SID (which may be deep in the packet), and then increment the offset field.

- o Even though the number of identifiers required within a controlled network is often less than 16 bit, and almost always 32 bits, carrying the overhead of 128 bits per SID in SRv6 can be seen as a significant unnecessary overhead, and workarounds such a proposed

micro programs [I-D.bonica-6man-comp-rtg-hdr], [I-D.bonica-spring-srv6-plus], [I-D.filsfils-spring-net-pgm-extension-srv6-usid] require complex forwarding plane processing and SRv6 programmability in the lower 64 bit is not required in the majority of use-cases for SIDs on midpoints.

For use-cases like this, it would be a lot easier to innovate IPv6 by clone & modify: E.g.: defining (say) IPv7 to be similar to IPv6, but without the constraints that are not useful for the controlled network use-case. A better alternative would be to create different profiles of IPv6 with [RFC8200] being one. However, there is, as yet, no concept of "profiles" in IPv6.

The issue of IP protocol operation in limited domains is discussed in [RFC8799].

Some possible solutions are described in [I-D.herbert-6man-eh-attrib]. This will be considered further in a future version of this text.

4.1.1.2. IPv6 for Edge-Compute

Today, the majority of end-to-end connections already do not pass via the traditional "Internet-Path" but instead toward a server in data center co-located with the access service provider Edge-2-Edge-EP [DOT]. In this case, there is no transit service provider, but there is a well-established commercial relationship between either end of the communications and the access service provider.

Today, the majority of traffic consists of video-streaming/TV services, but in the future, Edge-Compute will enable ever more applications to operate in such a controlled environment.

The difference between the aforementioned use-case of IPv6 within an service provider, and this use-case is that enhanced services in this would naturally operate end-to-end between a Data Center application server and the subscriber endpoints.

In the case of SRv6, it is not necessary to incur the overhead of an IPv6 in IPv6 encapsulation, the SRH can be inserted by the endpoint and removed by the endpoint on the other side. Nevertheless, the [RFC8200] limitations of not being able to add/remove or freely change the content of the SRH payload or any other EH on a midpoint router still exists. This seriously limits the usage and evolution of IPv6 to the edge-to-edge model.

4.1.1.3. Hop-by-Hop Extension Header processing

Hop-by-hop IPv6 extension headers caused interoperability and performance issues and as a result caused resistance to further leverage and extend them except for SRv6-SRH RPL-SRH [RFC6554]. In the authors opinions, this regression on hop-by-hop extension headers is because of a combination of insufficient specifications and resulting implementation issues. Both could be solved in future work with new hop-by-hop processing specifications.

For example, router alert (RA) was (and still maybe) implemented in routers so that all router alert packets are punted from the fast-path to the slow-path even when the "value" field identifies a protocol that the router can not process. As a result, protocols that rely on RA such as RSVP [RFC2205] or even more so Pragmatic General Multicast (PGM) [RFC3208] where filtered in networks because they caused high control plane load on routers that did not support either protocols but still unnecessarily punted their packets with RA.

There are no normative statements about the need that fast-path forwarding planes "MUST" be able to ignore unsupported/not-enabled EH features at a speed such that such a packet can be forward at the same speed as the same packet without the EH. For example, for RA, there is only a "SHOULD" requirement to do this in [RFC6398], a BCP published a decade after IPv6 router alert [RFC2711]. With such a gap in time between the specification and the BCP, it is impossible to rely on the existing RA and expect safe deployment across the Internet without still running into performance issues.

4.1.1.4. Segment Routing Header Constraints

The same design paradigm could have been used for the Segment Routing Header (SRH) [RFC8754], but there is no distinction possible for IPv6 instances running in such a controlled network or running as an Internetwork instance to form the Internet. This is particularly unfortunate as we are evolving to a model where, as noted earlier in this document, in most cases the packet will only travel through two well-known networks: the hosts network and the service provider network hosting the server to which the client is interacting.

4.1.2. Fixed Address Length

When IPv6 was designed, the key focus was on solving the problem of growth of the Internet and resulting growth of global Internet address space. Variable length and a heterogeneous address approach were proposed [RFC1347] however, these were rejected partially for

political reasons and partially out of a concern over the difficulty of parsing the packet and doing a fast address lookup.

There was seemingly no focus on better supporting the now millions of often network-layer isolated TCP/IP networks in industrial, defense, research, embedded, industrial or other commercial environments.

One key problems with with 128 bit addresses is the overhead on low-speed radio/IoT-wire networks. This is especially the case when using source-routing, where multiple of these addresses have to be included in the header. Current solutions are only able to resolve these issues with CPU expensive IETF standardized header compression techniques [RFC2507], [RFC3095], [RFC5795]. Even though these approaches are feasible in many of todays IoT networks, there is a strong desire to reduce power consumption in such devices. This is particularly the case where they are powered by a single-for-life-battery, or are self-powering through automatic replenished energy sources. As a result of this CPU performance in future IoT network should not be expected to increase but whenever feasible is more likely to decrease.

Another, often overlooked, problem of the 128 bit IPv6 addresses is that global address prefix allocation is a a big up-front burden on many IoT networks, but also isolated networks (industrial, defense, research, industrial). Often, this leads to the use of Unique Local Addresses (ULA) [RFC4193], which have the risk of conflicts when those previously isolated networks need to interconnect with other networks.

A further insight into the issues of IPv6 address lengths of 128 bits can be seen in the tussle over how to compress the address lengths in Segment Routing and network programming (in no particular order):

[I-D.bonica-6man-comp-rtg-hdr], [I-D.bonica-6man-crh-helper-opt], [I-D.bonica-spring-sr-mapped-six], [I-D.cheng-spring-shorter-srv6-sid-requirement], [I-D.decraene-spring-srv6-vlsid], [I-D.filsfilscheng-spring-srv6-srh-comp-sl-enc], [I-D.lc-6man-generalized-srh], [I-D.li-spring-compressed-srv6-np], [I-D.mirsky-spring-unified-id-network-programming], [I-D.steinberg-6man-crh-vs-sr-mpls], [I-D.templin-6man-crh-variable].

The root cause of this debate is the inflexibility of IPv6 in terms of its address length and semantics.

While solutions to these problems may look easier enough, it should be noted that in the time when IPv6 was designed, variable length addresses in the fastest forwarding planes were not seen as feasible,

and there was also a lack of experience with the impact of interconnecting heterogeneous address spaces other than as ships-in-the-night parallel operation of protocols. A lot of that experience came later through 14++ IPv4/IPv6 transition solutions designed in the past 20 years and respective work on address discovery in IETF frameworks such as SIP/STUN/ICE.

Another issue with the fixed length homogeneous address approach is the constraints this places on the current practice of overloading addresses with other functionality for example [RFC8986].

Since the original decision to only support fixed length packet addresses was taken there has been a significant improvement in the packet lookup capability of hardware. This is has been driven by the need to perform complex ACL lookup for security reasons and the interest in flow based techniques such as OpenFlow. It is thus worth revisiting the decision to only allow a single fixed address length and format.

4.2. Better Than Best Effort E2E Network Services

Some of the fastest growing network segments where new services are being introduced in an End-2-End manner belong to deployment models as described in [I-D.bryant-arch-fwd-layer-uc]. The requirements here for service delivery involves stringent E2E latency with no retransmission and no packet loss. Not all scenarios need "lower" latency but bounded to a particular value/range. Example use cases involving an user equipment (UE) consuming service from the provider cloud network or another UE (e.g. Vehicular device, IIoT) in the same network. Here the service endpoints could be connected over wire or wireless (LTE/NR) and the service termination happens in the provider network either close to the access network or provider core network. The existing network layer and best-effort model simply cannot guarantee needed service level objectives in these scenarios.

Some specific needs and requirements from cellular fixed transport networks are:

- o Need for determinism on E2E throughput and latency. The current TCP/IP is hence not-suitable for Mission-critical and real-time E2E applications.

- o Need for E2E QoS for ultra-reliable-low-latency communications (uRLLC).

- o Efficient use of protocols in the network by minimizing tunnels over tunnels and duplicate header fields.

o Efficient deployment of network slicing

4.3. Adaptive Bit-rate Video streaming

Even without going to future application requirements as described elsewhere in this document, even the majority of existing Internet traffic is lacking competitively usable and standardized service to support quality of service.

The majority of traffic today is Adaptive Bit-rate (ABR) based audio/video streaming. The primary benefit of this approach is that it can adjust itself to much lower bandwidth than the bandwidth to offer the ideal/target experience quality to the user. It therefore enabled Over The Top (OTT) services to offer streaming media. Nevertheless, ABR itself does not provide any actual quality guarantees.

Service providers that use ABR streaming to their subscribers do therefore combine ABR with IP developments, some non-published, which are often out-of-band bandwidth reservation schemes. These allow ABR video streams to have their ideal/target experience bandwidth within the SP's network and only need to degrade if there was bandwidth contention in the subscribers (home) network.

If a subscriber, or a content provider which is not the access service provider wanted to get the same type of bandwidth guarantees for other content across the access providers network, they could do so with existing IETF standards via RSVP [RFC2205] which is widely implemented, or NSIS [RFC4080], which was to the knowledge of the authors never implemented in widely used router products (because it does not offer sufficient benefits over RSVP). In either case, the per-flow control-plane based signaling architecture including the aforementioned router-alert issues make these protocols a difficult, likely not future-proof solution.

Even more fundamentally, ABR has shown that media streaming can easily support elastic adjustment between a range of bandwidth limits in which the quality is between acceptable and ideal, but there is as of today no standardized mechanisms by which to express relative bandwidth allocations when streams compete against each other that goes beyond the very loosely defined "internet fairness". For example, more intelligent congestion management could defend bandwidth the more the bandwidth approaches the minimum acceptable bandwidth, or admission control of bandwidth could be elastic. Some work in these direction exists in [RFC8698] with its ability for weighted congestion control or [I-D.ietf-tsvwg-intserv-multiple-tspec] for (limited) elastic admission control management.

4.4. Limited Domain Opportunities

Strictly of course this refers to the opportunities that the acceptance of limited domains [RFC8799] provides to the network operator in terms of the flexibility to enhance packet delivery in cases of high value traffic.

The removal of the constraint of a globally uniform protocol, such as unenhanced IPv6 would allow a best in class, domain specific forwarding layer to be deployed without the constant of the requirement that the protocol needed to serve all purposed, for all applications in all parts of the global network.

These opportunities are are further enhanced by noting that the delivery protocol to the application server, which as noted elsewhere in this text is moving closer to the edge, does not need to be the same as the host to application protocol since this is increasingly being opaquely tunneled over the delivery protocol. Furthermore, any distributed set of application servers maybe in their own domain, and this is not constrained to the same protocol that is used between the client and the server.

Clearly their are costs and complexities associated with moving from a globally heterogeneous protocol to a domain specific protocol, but the deciding factors are whether the application is deliverable over a globally general purpose forwarding layer, and whether there application and delivery system are economically attractive.

4.5. DetNet and Higher Precision Networking Service

Time Critical (TC), Ultra-Reliable, Low Latency (URLLC), Internet-of-Things is another important use case scenario-set that highlights requirements that are difficult to satisfy with existing Internet connectivity paths where a part of that path includes a radio access link. These kind of close-loop control systems borne over heterogeneous communications networks have very precision and bounded latency requirements for the E2E network connecting the sensor and actuator.

Deterministic networking within the IETF is focused on only one dimension of the URLLC problem.

DetNet is also far from attempting to identify currently if/how the services it plans to introduce could be made to operate over the Internet in general, instead, it focuses mostly on the shorter term goal to enable them in controlled networks within a limited domains.

Currently, the requirements for a DetNet forwarding plane have been reasonably mapped out for an MPLS based forwarding layer. Nevertheless, in addressing these needs within an IP network [RFC8939] the solution has of necessity been limited to the capabilities of the IP as it exists today. It has not, for example, been possible to add the packet replication elimination and reordering function (PREOF)which allows multiple concurrent packet delivery attempts in an MPLS network [RFC8964]. The DETNET body of requirements needs to be revisited in the light of any development to network forwarding capabilities.

4.6. Forwarding Plane vs. Control Plane

High-end hardware with accelerated forwarding plane devices, can support a significant number of forwarding states including destination entries (IP destination/mask, MPLS label, SR SID) as well as 2, 3 or 5 tuple IP/IPv6 "flow" entries. Nevertheless, the control plane that builds and changes these entries often limits their usability because the control plane does not even scale to the number of hardware accelerated forwarding entries possible, or because the supported rate of changes is slow.

The root of this problem is that with the increase of speed and scale of hardware accelerated forwarding hardware, control plane had challenges to keep up in performance. The performance of appropriately priced control plane CPUs (relative to the cost of the forwarding plane) has not grown at the same speed as that of hardware accelerated forwarding plane chips.

One of the directions to overcome these challenges is invisible outside these forwarder devices and it is to optimize the control-plane to forwarding plane interactions, such as programming the building of forwarding state directly on the accelerated forwarding infrastructure (e.g. NPU), but using otherwise existing control plane protocols.

A more fundamental approach is to redesign control plane protocols such that they are lighter weight in their signaling and state machinery, and can therefore be completely implemented in the hardware accelerated forwarding plane. Effectively turning a control plane protocol into an advanced forwarding plane protocol function.

This approach is logically most easily applicable to on-path per-flow signaling mechanisms such as RSVP or RSVP-TE, both of which are quite complex with their signaling messaging and state keeping and therefore directly infeasible to become hardware accelerated forwarding implementations. An example approach to provide similar functionality to RSVP with signaling light-weight enough to allow

hardware accelerated implementation are the in-band signaling mechanisms (e.g. for TCP or UDP) described in [DIP1] [I-D.han-tsvwg-ip-transport-qos] [I-D.han-tsvwg-enhanced-diffserv].

Signaling that is feasible to become part of a complete in-forwarding-plane signaling solution is not limited to in-band on-path flow signaling, but would likely also be applied to other signaling options. Of the aforementioned existing signaling protocols, IGPs are likely the ones whose signaling could most easily be processed in an NPU compute elements except that the SPF calculation itself introduces a complexity that would make this very complex. One example of a solution that solves this problem by signaling the actual per-hop adjacencies in IGP and therefore eases NPU implementation can be found in [I-D.chunduri-isis-preferred-path-routing].

In summary: The scope of what should be considered forwarding plane today is defined by decade historic architectures, but should for the future be scoped by the realities of the new, different "layers" of hardware and their capabilities. Hence also the use of the term forwarding plane, because it can span not only across classical bridging (L2), label/tag/SIG switching (L2.5), network/internetwork (L3) and transport (L4) layers, but also across the classical "data plane" and "control plane" components of each such layer.

4.7. User-Network/Network-User Interface Signaling

Some of the deployment models as described in [I-D.bryant-arch-fwd-layer-uc], needs specific signaling mechanism from user/applications. These are needed for E2E service offering for better than best effort Section 4.2 or high-precision networking Section 4.5. These may involve new transport mechanisms at hosts, middle-boxes and routers to meet the E2E service requirements in these limited domain deployments.

Here one of the functional requirements is to signal the service level objectives (SLOs) dynamically for a particular service from the network. This signaling includes the service description, the service negotiation with the network, the service setup or modification, or the need to execute some functions at network device and send the results back to the sender. However, the current IP was not designed for this. For example, the result of SLO negotiation at any hop needs to be updated in the IP packet at the router and returned back to the sender (originating host or gateway device for a Service Provider).

There are some attempts to achieve the above as described in [I-D.han-tsvwg-ip-transport-qos], which describes general in-band

signaling for QoS control with IPv6 protocol and [I-D.han-tsvwg-enhanced-diffserv], which proposes a backward compatible class-based queuing and scheduling schema for hybrid service to support guaranteed service from the network (e.g. for latency and bandwidth).

In summary, it is difficult to do better than best effort or High Precision Services described in Section Section 4.5, in closed domains with current IP given the best effort congestion control (TCP/QUIC) and explicit congestion notification (ECN) framework. A comprehensive mechanism needs to be explored as the limitations in silicon technologies or deployment models 30 years ago are not relevant with respect to security, scalability, packet size change, MSS or FCS recalculation, etc.

5. Candidate Solution Directions

This section is an incomplete list of solution considerations, but is not prescriptive about any specific approach or technical solution, and is provided to stimulate thought on the subject.

5.1. Variable Length Addresses

When private networks are set up, they only need to use an address length that allows the construction of networks sufficiently large to meet the expected service requirements. If a future network layer protocol could support address length of e.g.: 16, 24, 32, 48, 64 and 128 bits (or maybe more), it would be easy for such networks to pick a right size. This would allow them to have as efficient packets without compression as possible, and it would also avoid for them to have to think about allocation procedures for "global" addresses.

Whenever networks with a smaller address size would later on have to interconnect to other networks, the shorter length address would have to be interpreted as the suffix of a sufficiently larger address space through which those connecting networks could achieve unique, non-overlapping addresses. At the border between these networks, high speed forwarding planes could easily perform per-packet stateless prefix addition/deletion transformations of addresses in the packet header when the interconnection should be free of further policy. When such an interconnection is desired to employ specific traffic control policies, mapping of addresses in a stateful manner is a convenient way to enforce and support such policies through the forwarding plane.

5.2. Address Semantics

Classically IP unicast addresses identify an interface. There is the special case of a loop-back address, but this is normally modeled as an internal interface. Addresses are often silently mapped to include other semantics and this is most developed in the IP network programming concept [RFC8986].

MPLS is more general. It defines the concept of a Forwarding Equivalence Class in which a Label which can be visualized as an offset into a specific table with up to 2^20 entries, with the table containing the instruction to be executed. Thus a single identifier is able to specify: forward towards an egress, forward along a specific path, decapsulate and sent to an interface, decapsulate and forward via an IP lookup in a label specific address table etc.

The semantics of the MPLS label and the size of the label are such that it is not possible to include any instruction parameters in the label and very inefficient to include those parameters in one or more further labels. The only example of doing this is the Entropy Label indicator [RFC6790] which uses two Label Stack Entries (LSEs). Any future development along these lines will need at least three LSEs.

Whilst an IPv6 is larger there is still limited space to add parameters within the address. In the current work on this the size is limited to 16 bits, and there is a fundamental limit of 64 bits.

It is clear that move is towards a multiplicity of semantic for the network layer address, and indeed a formal recognition that the address is in reality an instruction with a specific scope.

5.3. Multiple Instructions

What we have learned from MPLS and then from SRv6 is that it is often desirable for a node (be that the originating host or a router) to impose on a packet a set of instructions to be executed in sequence by one or more entities in the network. An development of IP or any successor needs to recognize this and provide a simple and efficient way to incorporate a list (or stack) of instructions within the packet header.

5.4. Node and Path Specific Processing Instructions

There is an established need to do node specific instructions as is indicated by the design of MPLS and Segment Routing (SR). Any development of the forwarding system needs to retain this feature and ideally develop a method that is simultaneously both general and efficient.

References to efficiency include efficiency in packet size and efficiency decoding and and executing the instruction. The efficiency of encoding is not simply a matter of on the wire bandwidth, but is also a matter of the size of the forwarder packet header cache. This cache has to operate at wire speed can be an expensive silicon element.

There is also a need to do path specific operations as are done in RSVP-TE. However RSVP has a significant path set-up and path maintenance cost. Clearly a per path instruction can be specified as a set of N per node instructions where N is the number of hops along the path, for example by using SR, but that is not an efficient encoding where N is large. It is thus a useful optimization to include the ability to include per path instructions, and this is the subject of further study.

5.5. Integrated Assurance and Verification

Being best effort in nature, assurance for services provided using IP is left to add-on solutions built after the fact. How to perform tasks such as verifying of service levels is left as an exercise for network providers, often approached using statistical approaches that are themselves "best effort" in nature. This will be no longer sufficient for mission-critical services such as tele-driving or tele-operations that demand guarantees, where failure to meet those guarantees may expose providers and users exposed to liability demands and call the feasibility of applications relying on those services into question.

Moving forward, network protocols suitable to deliver high-precision services for mission critical applications need to address assurance as an intrinsic property, not left to afterthoughts.

5.6. For Consideration in a Future Version

A future version of this document will consider E2E communication beyond-best-effort, high precision services, high precision telemetry, E2E Volumetric data transfer and high precision congestion control beyond that provided by the diffserv QoS bits.

6. IANA Considerations

This document does not request any allocations from IANA.

7. Security Considerations

Security is likely to be more significant with the applications being considered in this work. With interest in tightly controlled access and latency, and contractual terms of business it is going to be necessary to have provable right of access to network resources. However heavyweight security is a contra-requirement to the light-weight process needed for power efficiency, fast forwarding and low latency. Addressing this will require new insights into network security.

Further information on the issue of providing security in latency sensitive environments can be found in [RFC9055] which are a sub-set of the considerations applicable to the new use cases considered in this text.

8. References

8.1. Normative References

[RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, DOI 10.17487/RFC0791, September 1981, <https://www.rfc-editor.org/info/rfc791>.

[RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <https://www.rfc-editor.org/info/rfc8200>.

8.2. Informative References

[DIP1] ETSI, "Recommendation for New Transport Technologies, GR NGP 010", September 2018, <https://www.etsi.org/deliver/etsi_gr/NGP/001_099/010/01.01.01_60/gr_NGP010v010101p.pdf>.

[DOT] Huston, G., "The Death of Transit and Beyond", n.d., <https://hknog.net/wp-content/uploads/2018/03/01_GeoffHuston_TheDeath_of_Transit_and_Beyond.pdf>.

[I-D.bonica-6man-comp-rtg-hdr] Bonica, R., Kamite, Y., Alston, A., Henriques, D., and L. Jalil, "The IPv6 Compact Routing Header (CRH)", draft-bonica-6man-comp-rtg-hdr-27 (work in progress), November 2021.

[I-D.bonica-6man-crh-helper-opt]
Li, X., Bao, C., Ruan, E., and R. Bonica, "Compressed Routing Header (CRH) Helper Option", draft-bonica-6man-crh-helper-opt-04 (work in progress), October 2021.

[I-D.bonica-spring-sr-mapped-six]
Bonica, R., Hegde, S., Kamite, Y., Alston, A., Henriques, D., Jalil, L., Halpern, J., Linkova, J., and G. Chen, "Segment Routing Mapped To IPv6 (SRm6)", draft-bonica-spring-sr-mapped-six-04 (work in progress), September 2021.

[I-D.bonica-spring-srv6-plus]
Bonica, R., Hegde, S., Kamite, Y., Alston, A., Henriques, D., Jalil, L., Halpern, J., Linkova, J., and G. Chen, "Segment Routing Mapped To IPv6 (SRm6)", draft-bonica-spring-srv6-plus-06 (work in progress), October 2019.

[I-D.bryant-arch-fwd-layer-uc]
Bryant, S., Chunduri, U., Eckert, T., and A. Clemm, "Forwarding Layer Use Cases", draft-bryant-arch-fwd-layer-uc-03 (work in progress), January 2022.

[I-D.cheng-spring-shorter-srv6-sid-requirement]
Cheng, W., Chongfeng, Pang, R., Li, Z., Chen, R., Lijun, Duan, X., Mirsk, G., Dukes, D., and S. Zadok, "Shorter SRv6 SID Requirements", draft-cheng-spring-shorter-srv6-sid-requirement-02 (work in progress), July 2020.

[I-D.chunduri-isis-preferred-path-routing]
Chunduri, U., Li, R., White, R., Tantsura, J., Contreras, L. M., and Y. Qu, "Preferred Path Routing (PPR) in IS-IS", draft-chunduri-isis-preferred-path-routing-00 (work in progress), June 2018.

[I-D.decraene-spring-srv6-vlsid]
Decraene, B., Raszuk, R., Li, Z., and C. Li, "SRv6 vSID: Network Programming extension for variable length SIDs", draft-decraene-spring-srv6-vlsid-06 (work in progress), September 2021.

[I-D.filsfils-spring-net-pgm-extension-srv6-usid]
Filsfils, C., Garvia, P. C., Cai, D., Voyer, D., Meilik, I., Patel, K., Henderickx, W., Jonnalagadda, P., Melman, D., Liu, Y., and J. Guichard, "Network Programming extension: SRv6 uSID instruction", draft-filsfils-spring-net-pgm-extension-srv6-usid-12 (work in progress), December 2021.

[I-D.filsfilscheng-spring-srv6-srh-comp-sl-enc]
Cheng, W., Filsfils, C., Li, Z., Cai, D., Voyer, D., Clad, F., Zadok, S., Guichard, J. N., and L. Aihua, "Compressed SRv6 Segment List Encoding in SRH", draft-filsfilscheng-spring-srv6-srh-comp-sl-enc-03 (work in progress), May 2021.

[I-D.han-tsvwg-enhanced-diffserv]
Han, L., Qu, Y., and R. Li, "Enhanced DiffServ by In-band Signaling", draft-han-tsvwg-enhanced-diffserv-00 (work in progress), November 2019.

[I-D.han-tsvwg-ip-transport-qos]
Han, L., Qu, Y., Dong, L., Li, R., Nadeau, T., Smith, K., and J. Tantsura, "Resource Reservation Protocol for IP Transport QoS", draft-han-tsvwg-ip-transport-qos-03 (work in progress), October 2019.

[I-D.herbert-6man-eh-attrib]
Herbert, T., "Attribution Option for Extension Header Insertion", draft-herbert-6man-eh-attrib-03 (work in progress), October 2020.

[I-D.ietf-ippm-ioam-data]
Brockners, F., Bhandari, S., and T. Mizrahi, "Data Fields for In-situ OAM", draft-ietf-ippm-ioam-data-17 (work in progress), December 2021.

[I-D.ietf-tsvwg-intserv-multiple-tspec]
Polk, J. and S. Dhesikan, "Integrated Services (IntServ) Extension to Allow Signaling of Multiple Traffic Specifications and Multiple Flow Specifications in RSVPv1", draft-ietf-tsvwg-intserv-multiple-tspec-02 (work in progress), February 2013.

[I-D.lc-6man-generalized-srh]
Li, Z., Li, C., Cheng, W., Xie, C., Li, C., Tian, H., and F. Zhao, "Generalized Segment Routing Header", draft-lc-6man-generalized-srh-03 (work in progress), February 2021.

[I-D.li-spring-compressed-srv6-np]
Li, Z., Li, C., Xie, C., LEE, K., Tian, H., Zhao, F., Guichard, J. N., Li, C., and S. Peng, "Compressed SRv6 Network Programming", draft-li-spring-compressed-srv6-np-02 (work in progress), February 2020.

[I-D.mirsky-spring-unified-id-network-programming]
Weiqiang, C., Mirsky, G., Aihua, L., and P. Shaofu, "SRv6 network programming using Unified Identifier", draft-mirsky-spring-unified-id-network-programming-00 (work in progress), March 2020.

[I-D.steinberg-6man-crh-vs-sr-mpls]
Steinberg, D., Henderickx, W., Li, Z., Cheng, W., and D. Voyer, "SR-MPLS over IPv6 satisfies CRH requirements", draft-steinberg-6man-crh-vs-sr-mpls-00 (work in progress), June 2020.

[I-D.templin-6man-crh-variable]
Templin, F. L., "IPv6 Compressed Routing Header with Variable Length Addresses", draft-templin-6man-crh-variable-00 (work in progress), May 2020.

[RFC1347] Callon, R., "TCP and UDP with Bigger Addresses (TUBA), A Simple Proposal for Internet Addressing and Routing", RFC 1347, DOI 10.17487/RFC1347, June 1992, <https://www.rfc-editor.org/info/rfc1347>.

[RFC2205] Braden, R., Ed., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, DOI 10.17487/RFC2205, September 1997, <https://www.rfc-editor.org/info/rfc2205>.

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, DOI 10.17487/RFC2460, December 1998, <https://www.rfc-editor.org/info/rfc2460>.

[RFC2507] Degermark, M., Nordgren, B., and S. Pink, "IP Header Compression", RFC 2507, DOI 10.17487/RFC2507, February 1999, <https://www.rfc-editor.org/info/rfc2507>.

[RFC2711] Partridge, C. and A. Jackson, "IPv6 Router Alert Option", RFC 2711, DOI 10.17487/RFC2711, October 1999, <https://www.rfc-editor.org/info/rfc2711>.

[RFC3095] Bormann, C., Burmeister, C., Degermark, M., Fukushima, H., Hannu, H., Jonsson, L-E., Hakenberg, R., Koren, T., Le, K., Liu, Z., Martensson, A., Miyazaki, A., Svanbro, K., Wiebke, T., Yoshimura, T., and H. Zheng, "RObust Header Compression (ROHC): Framework and four profiles: RTP, UDP, ESP, and uncompressed", RFC 3095, DOI 10.17487/RFC3095, July 2001, <https://www.rfc-editor.org/info/rfc3095>.

[RFC3208] Speakman, T., Crowcroft, J., Gemmell, J., Farinacci, D., Lin, S., Leshchiner, D., Luby, M., Montgomery, T., Rizzo, L., Tweedly, A., Bhaskar, N., Edmonstone, R., Sumanasekera, R., and L. Vicisano, "PGM Reliable Transport Protocol Specification", RFC 3208, DOI 10.17487/RFC3208, December 2001, <https://www.rfc-editor.org/info/rfc3208>.

[RFC4080] Hancock, R., Karagiannis, G., Loughney, J., and S. Van den Bosch, "Next Steps in Signaling (NSIS): Framework", RFC 4080, DOI 10.17487/RFC4080, June 2005, <https://www.rfc-editor.org/info/rfc4080>.

[RFC4193] Hinden, R. and B. Haberman, "Unique Local IPv6 Unicast Addresses", RFC 4193, DOI 10.17487/RFC4193, October 2005, <https://www.rfc-editor.org/info/rfc4193>.

[RFC4302] Kent, S., "IP Authentication Header", RFC 4302, DOI 10.17487/RFC4302, December 2005, <https://www.rfc-editor.org/info/rfc4302>.

[RFC5795] Sandlund, K., Pelletier, G., and L-E. Jonsson, "The RObust Header Compression (ROHC) Framework", RFC 5795, DOI 10.17487/RFC5795, March 2010, <https://www.rfc-editor.org/info/rfc5795>.

[RFC6282] Hui, J., Ed. and P. Thubert, "Compression Format for IPv6 Datagrams over IEEE 802.15.4-Based Networks", RFC 6282, DOI 10.17487/RFC6282, September 2011, <https://www.rfc-editor.org/info/rfc6282>.

[RFC6398] Le Faucheur, F., Ed., "IP Router Alert Considerations and Usage", BCP 168, RFC 6398, DOI 10.17487/RFC6398, October 2011, <https://www.rfc-editor.org/info/rfc6398>.

[RFC6554] Hui, J., Vasseur, JP., Culler, D., and V. Manral, "An IPv6 Routing Header for Source Routes with the Routing Protocol for Low-Power and Lossy Networks (RPL)", RFC 6554, DOI 10.17487/RFC6554, March 2012, <https://www.rfc-editor.org/info/rfc6554>.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <https://www.rfc-editor.org/info/rfc6790>.

[RFC8655] Finn, N., Thubert, P., Varga, B., and J. Farkas, "Deterministic Networking Architecture", RFC 8655, DOI 10.17487/RFC8655, October 2019, <https://www.rfc-editor.org/info/rfc8655>.

[RFC8698] Zhu, X., Pan, R., Ramalho, M., and S. Mena, "Network-Assisted Dynamic Adaptation (NADA): A Unified Congestion Control Scheme for Real-Time Media", RFC 8698, DOI 10.17487/RFC8698, February 2020, <https://www.rfc-editor.org/info/rfc8698>.

[RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <https://www.rfc-editor.org/info/rfc8754>.

[RFC8799] Carpenter, B. and B. Liu, "Limited Domains and Internet Protocols", RFC 8799, DOI 10.17487/RFC8799, July 2020, <https://www.rfc-editor.org/info/rfc8799>.

[RFC8939] Varga, B., Ed., Farkas, J., Berger, L., Fedyk, D., and S. Bryant, "Deterministic Networking (DetNet) Data Plane: IP", RFC 8939, DOI 10.17487/RFC8939, November 2020, <https://www.rfc-editor.org/info/rfc8939>.

[RFC8964] Varga, B., Ed., Farkas, J., Berger, L., Malis, A., Bryant, S., and J. Korhonen, "Deterministic Networking (DetNet) Data Plane: MPLS", RFC 8964, DOI 10.17487/RFC8964, January 2021, <https://www.rfc-editor.org/info/rfc8964>.

[RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/info/rfc8986>.

[RFC9055] Grossman, E., Ed., Mizrahi, T., and A. Hacker, "Deterministic Networking (DetNet) Security Considerations", RFC 9055, DOI 10.17487/RFC9055, June 2021, <https://www.rfc-editor.org/info/rfc9055>.

Authors' Addresses

Stewart Bryant
University of Surrey 5/6GIC

Email: sb@stewartbryant.com

Uma Chunduri
Intel

Email: umac.ietf@gmail.com

Toerless Eckert
Futurewei Technologies Inc.

Email: tte@cs.fau.de

Alexander Clemm
Futurewei Technologies Inc.

Email: ludwig@clemm.org