DetNet
Internet-Draft
Intended status: Standards Track
Expires: September 4, 2016

N. Finn
P. Thubert
Cisco
M. Johas Teener
Broadcom
March 3, 2016

Deterministic Networking Architecture
draft-finn-detnet-architecture-03

Abstract

Deterministic Networking (DetNet) provides a capability to carry specified unicast or multicast data flows for real-time applications with extremely low data loss rates and bounded latency. Techniques used include: 1) reserving data plane resources for individual (or aggregated) DetNet flows in some or all of the relay systems (bridges or routers) along the path of the flow; 2) providing fixed paths for DetNet flows that do not rapidly change with the network topology; and 3) sequentializing, replicating, and eliminating duplicate packets at various points to ensure the availability of at least one path. The capabilities can be managed by configuration, or by manual or automatic network management.

Table of Contents

1. Introduction

Deterministic Networking (DetNet) is a service that can be offered by a network to data flows (DetNet flows) that that are limited, at their source, to a maximum data rate specified by that source. DetNet provides these flows extremely low packet loss rates and assured maximum end-to-end delivery latency. This is accomplished by dedicating network resources such as link bandwidth and buffer space to DetNet flows and/or classes of DetNet flows. Unused reserved resources are available to non-DetNet packets.

The Deterministic Networking Problem Statement [I-D.finn-detnet-problem-statement] introduces Deterministic Networking, and Deterministic Networking Use Cases [I-D.grossman-detnet-use-cases] summarizes the need for it.

A goal of DetNet is a converged network in all respects. That is, the presence of DetNet flows does not preclude non-DetNet flows, and the benefits offered DetNet flows should not, except in extreme cases, prevent existing QoS mechanisms from operating in a normal fashion, subject to the bandwidth required for the DetNet flows. A single source-destination pair can trade both DetNet and non-DetNet flows. End systems and applications need not instantiate special interfaces for DetNet flows. Networks are not restricted to certain topologies; connectivity is not restricted. Any application that generates a data flow that can be usefully characterized as having a maximum bandwidth should be able to take advantage of DetNet, as long as the necessary resources can be reserved. Reservations can be made by the application itself, via network management, by an applications controller, or by other means.

Many applications of interest to Deterministic Networking require the ability to synchronize the clocks in end systems to a sub-microsecond accuracy. Some of the queue control techniques defined in Section 4.4 also require time synchronization among relay systems. The means used to achieve time synchronization are not addressed in this document.

The present document is an individual contribution, intended by the authors for eventual adoption by the DetNet working group. As such, it expresses the only the opinions of the authors.

2. Terminology

The following special terms are used in this document in order to avoid the assumption that a given element in the architecture does or does not have Internet Protocol stack, functions as a router, bridge, firewall, or otherwise plays a particular role at Layer-2 or higher. This section also serves as a dictionary for translating between IEEE 802 and DetNet terminology.

destination
: An end system capable of sinking a DetNet flow.

DetNet flow
: A DetNet flow is a sequence of packets from a single source, through some number of relay systems to one or more destinations, that is limited by the source in its maximum packet size and transmission rate, and can thus be ensured the DetNet Quality of Service (QoS) from the network.

end system
: Commonly called a "host" in IETF documents, and an "end station" is IEEE 802 documents. End systems of interest to this document are either sources or destinations.

listener
: The IEEE 802 term for a destination of a DetNet flow.

relay system
: A router, bridge, Label Switch Router (LSR), firewall, or any other system that forwards packets from one interface to another.

reservation
: A trail of configuration from source to destination(s) through relay systems associated with a DetNet flow, required to deliver the benefits of DetNet.

source
: An end system capable of sourcing a DetNet flow.

stream
: The IEEE 802 term for a DetNet flow.

talker
: The IEEE 802 term for the source of a DetNet flow.

3. Providing the DetNet Quality of Service

DetNet Quality of Service is expressed in terms of:

- Minimum and maximum end-to-end latency from source to destination;
- Probability of loss of a packet, assuming the normal operation of the relay systems and links;
- Probability of loss of a packet in the event of the failure of a relay system or link.

It is a distinction of DetNet that it is concerned solely with worst-case values for all of the above parameters. Average, mean, or typical values are of no interest, because they do not affect the ability of a real-time system to perform its tasks. For example, in this document, we will often speak of assuring a DetNet flow a bounded latency. In general, a trivial priority-based queuing scheme will give better average latency to a data flow than DetNet, but of course, the worst-case latency is essentially unbounded.

Three techniques are employed by DetNet to achieve these QoS parameters:

a. Zero congestion loss (Section 3.1). Network resources such as link bandwidth, buffers, queues, shapers, and scheduled input/output slots are assigned in each relay system to the use of a specific DetNet flow or class of DetNet flows. Given a finite amount of buffer space, zero congestion loss necessarily ensures a bounded end-to-end latency. Depending on the resources employed, a minimum latency, and thus bounded jitter, can also be achieved.

b. Pinned paths (Section 3.2). Point-to-point paths or point-to-multipoint trees through the network from a source to one or more destinations can be established, and DetNet flows assigned to follow a particular path or tree.

c. Packet replication and deletion (Section 3.3). End systems and/or relay systems can number packets sequentially, replicate them, and later eliminate all but one of the replicants, at multiple points in the network in order to ensure that one (or more) equipment failure events still leave at least one path intact for a DetNet flow.

These three techniques can be applied independently, giving eight possible combinations, including none (no DetNet), although some combinations are of wider utility than others. This separation keeps

the protocol stack coherent and maximizes interoperability with existing and developing standards in this (IETF) and other Standards Development Organizations. Some examples of typical expected combinations:

- Pinned paths (a) plus packet replication (b) are exactly the techniques employed by [HSR-PRP]. Pinned paths are achieved by limiting the physical topology of the network, and the sequentialization, replication, and duplicate elimination are facilitated by packet tags added at the front or the end of Ethernet frames.

- Zero congestion loss (a) alone is is offered by IEEE 802.1 Audio Video bridging [IEEE802.1BA-2011]. As long as the network suffers no failures, zero congestion loss can be achieved through the use of a reservation protocol (MSRP), shapers in every relay system (bridge), and a bit of network calculus.

- Using all three together gives maximum protection.

There are, of course, simpler methods available (and employed, today) to achieve levels of latency and packet loss that are satisfactory for many applications. Prioritization and over-provisioning is one such technique. However, these methods generally work best in the absence of any significant amount of non-critical traffic in the network (if, indeed, such traffic is supported at all), or work only if the critical traffic constitutes only a small portion of the network's theoretical capacity, or work only if all systems are functioning properly, or in the absence of actions by end systems that disrupt the network's operations.

There are any number of methods in use, defined, or in progress for accomplishing each of the above techniques. It is expected that this DetNet Architecture will assist various vendors, users, and/or "vertical" Standards Development Organizations (dedicated to a single industry) to make selections among the available means of implementing DetNet networks.

3.1. Zero Congestion Loss

The primary means by which DetNet achieves its QoS assurances is to completely eliminate congestion at an output port as a cause of packet loss. Given that a DetNet flow cannot be throttled, this can be achieved only by the provision of sufficient buffer storage at each hop through the network to ensure that no packets are dropped due to a lack of buffer storage.

Ensuring adequate buffering requires, in turn, that the source, and every relay system along the path to the destination (or nearly every relay system -- see Section 4.3.2) be careful to regulate its output to not exceed the data rate for any DetNet flow, except for brief periods when making up for interfering traffic. Any packet sent ahead of its time potentially adds to the number of buffers required by the next hop, and may thus exceed the resources allocated for a particular DetNet flow.

The low-level mechanisms described in Section 4.4 provide the necessary regulation of transmissions by an edge system or relay system to ensure zero congestion loss. The reservation of the bandwidth and buffers for a DetNet flow requires the provisioning described in Section 4.9.

3.2. Pinned paths

In networks controlled by typical peer-to-peer protocols such as IEEE 802.1 ISIS bridged networks or IETF OSPF routed networks, a network topology event in one part of the network can impact, at least briefly, the delivery of data in parts of the network remote from the failure or recovery event. Thus, even redundant paths through a network, if controlled by the typical peer-to-peer protocols, do not eliminate the chances of brief losses of contact.

Many real-time networks rely on physical rings or chains of two-port devices, with a relatively simple ring control protocol. This supports redundant paths with a minimum of wiring. As an additional benefit, ring topologies can often utilize different topology management protocols than those used for a mesh network, with a consequent reduction in the response time to topology changes. Of course, this comes at some cost in terms of increased hop count, and thus latency, for the typical path.

In order to get the advantages of low hop count and still ensure against even very brief losses of connectivity, DetNet employs pinned paths, where the path taken by a given DetNet flow does not change, at least immediately, and likely not at all, in response to network topology events. When combined with seamless redundancy (Section 3.3), this results in a high likelihood of continuous connectivity.

3.3. Seamless Redundancy

After congestion loss has been eliminated, the most important causes of packet loss are random media and/or memory faults, and equipment failures.

Seamless redundancy involves three capabilities:

- Adding sequence numbers, once, to the packets of a DetNet flow.
- Replicating these packets and, typically, sending them along at least two different paths to the destination(s). (Often, the pinned paths of Section 3.2.)
- Discarding duplicated packets.

In the simplest case, this amounts to replicating each packet in a source that has two interfaces, and conveying them through the network, along separate paths, to the similarly dual-homed destinations, that discard the extras. This ensures that one path (with zero congestion loss) remains, even if some relay system fails.

Alternatively, relay systems in the network can provide replication and elimination facilities at various points in the network, so that multiple failures can be accommodated.

This is shown in the following figure, where the two relay systems each replicate (R) the DetNet flow on input, sending the DetNet flow to both the other relay system and to the end system, and eliminate duplicates (E) on the output interface to the right-hand end system. Any one link in the network can fail, and the Detnet flow can still get through. Furthermore, two links can fail, as long as they are in different segments of the network.

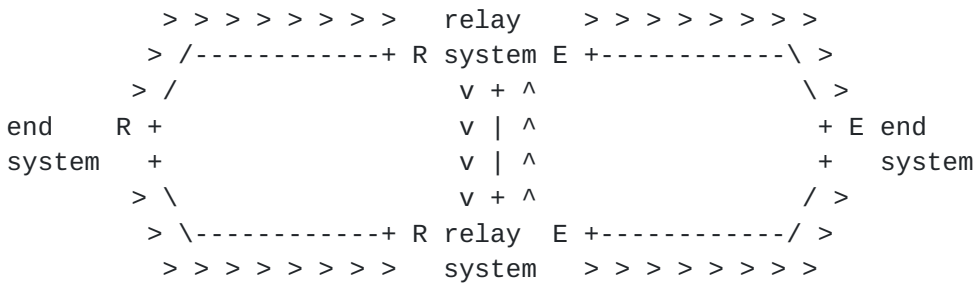

Figure 1

Note that seamless redundancy does not react to and correct failures; it is entirely passive. Thus, intermittent failures, mistakenly created access control lists, or misrouted data is handled just the same as the equipment failures that are detected handled by typical routing and bridging protocols.

4. DetNet Architecture

4.1. Elements of DetNet Architecture

The DetNet architecture has a number of elements, discussed in the following sections. Note that not every application requires all of these elements.

a. A model for the definition, identification, and operation of DetNet flows (Section 4.3), for use by relay systems to classify and process individual packets following per-flow rules.

b. A model for the flow of data out of an end system or through a relay system that can be used to predict the bounds for that system's impact on the QoS of a DetNet flow, for use by the Controllers to configure policing and shaping engines in Network Systems over the Southbound interface. The model includes:

    1. A model for queuing, transmission selection, shaping, preemption, and timing resources that can be used by an end system or relay system to control the selection of packets output on an interface. These models must have sufficiently well-defined characteristics, both individually and in the aggregate, to give predictable results for the QoS for DetNet packets (Section 4.4).

    2. A model for identifying misbehaving DetNet flows and mitigating their impact on properly functioning DetNet flows (Section 4.6).

c. A model for the relay system to inform the controller(s) of the information it needs for adequate path computations (Section 4.2) including:

    1. Systems' individual capabilities (e.g. can do replication, can do precise time).

    2. Link capabilities and resources (e.g. bandwidth, transmission delay, hardware deterministic support to the physical layer, ...)

    3. Physical resources (total and available buffers, timers, queues, etc)

    4. Network Adjacencies (neighbors)

d. A model for the provision of a service, by end systems or relay systems, to replicate and forward a DetNet flow over redundant paths. The model includes:

   1. A model for specifying multiple stable paths across a network that can perform packet forwarding at both Layer 3 and at lower layers, to which specific DetNet flows can be assigned (Section 4.2).

   2. A model and data plane format(s) for sequencing and replicating the packets of a DetNet flow, typically at or near the source, sending the replicated DetNet flows over different stable paths, merging and/or re-replicating those packets at other points in the network, and finally eliminating the duplicates, typically at or near the destination(s), in order to provide high availability (Section 3.3).

e. The protocol stack model for an end system and/or a relay system should support the above elements in a manner that maximizes the applicability of existing standards and protocols to the DetNet problem, and allows for the creation of new protocols only where needed, thus making DetNet an add-on feature to existing networks, rather than a new way to do networking. In particular this protocol stack supports networks in which the path from source to destination(s) includes bridges and/or routers in any order (Section 4.7).

f. A variety of models for the provisioning of DetNet flows can be envisioned, including orchestration by a central controller or by a federation of controllers, provisioning by relay systems and end systems sharing peer-to-peer protocols, by off-line configuration, or by a combination of these methods. The provisioning models are similar to existing Layer-2 and Layer-3 models, in order to minimize the amount of innovation required in this area (Section 4.9).

4.2. Traffic Engineering for DetNet

Traffic Engineering Architecture and Signaling (TEAS) [TEAS] defines traffic-engineering architectures for generic applicability across packet and non-packet networks. From TEAS perspective, Traffic Engineering (TE) refers to techniques that enable operators to control how specific traffic flows are treated within their networks.

Because if its very nature of establishing pinned optimized paths, Deterministic Networking can be seen as a new, specialized branch of

Traffic Engineering, and inherits its architecture with a separation into planes.

The Deterministic Networking architecture is thus composed of three planes, a (User) Application Plane, a Controller Plane, and a Network Plane, which echoes that of Software-Defined Networking (SDN): Layers and Architecture Terminology [RFC7426] which is represented below:

SDN Layers and Architecture Terminology per RFC 7426

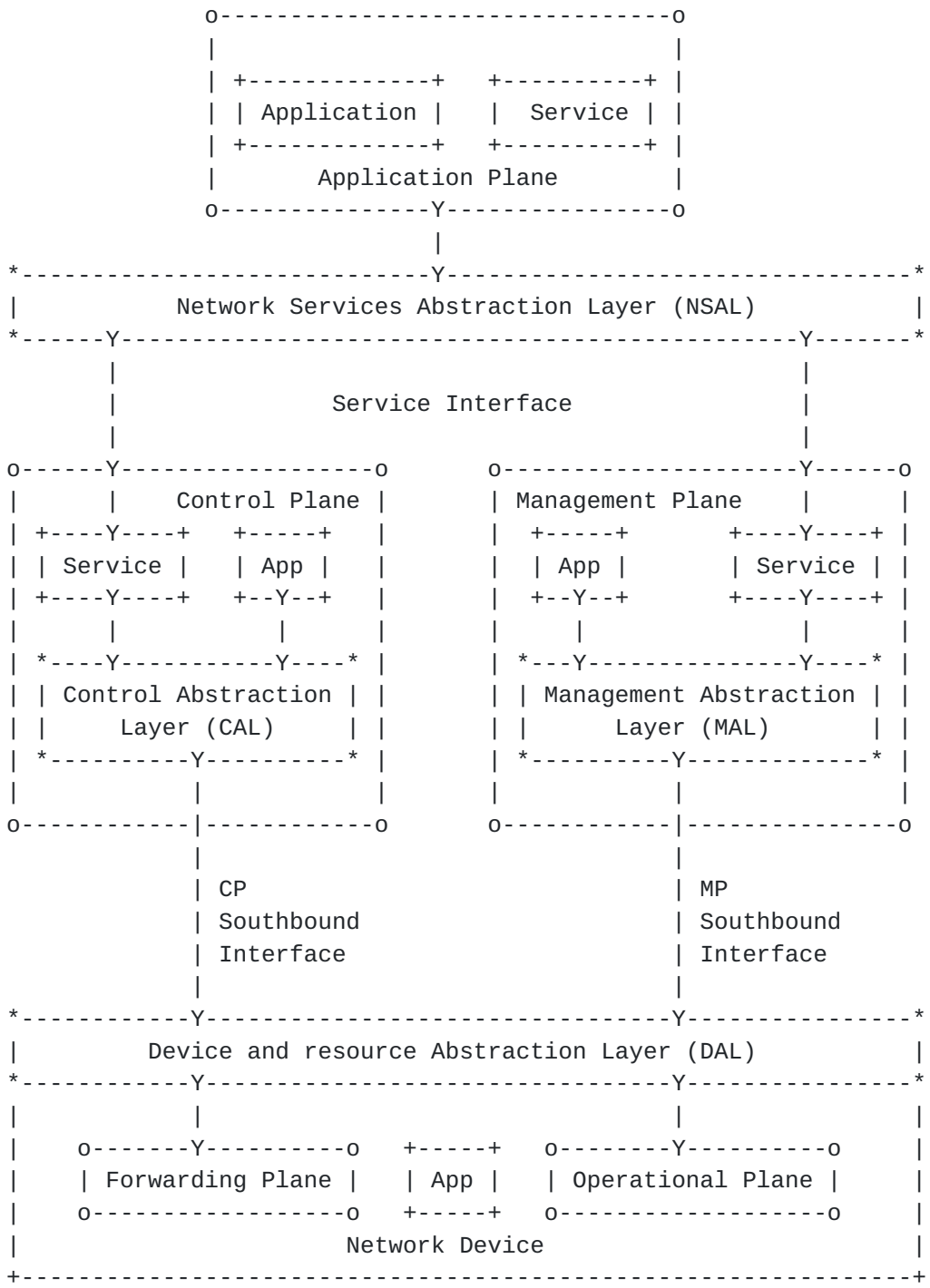

Figure 2

4.2.1. The Application Plane

Per [RFC7426], the Application Plane includes both applications and services. In particular, the Application Plane incorporates the User Agent, a specialized application that interacts with the end user / operator and performs requests for Deterministic Networking services via an abstract Flow Management Entity, (FME) which may or may not be collocated with (one of) the end systems.

At the Application Plane, a management interface enables the negotiation of flows between end systems. An abstraction of the flow called a Traffic Specification (TSpec) provides the representation. This abstraction is used to place a reservation over the (Northbound) Service Interface and within the Application plane. It is associated with an abstraction of location, such as IP addresses and DNS names, to identify the end systems and eventually specify intermediate relay systems.

4.2.2. The Controller Plane

The Controller Plane corresponds to the aggregation of the Control and Management Planes in [RFC7426], though Common Control and Measurement Plane (CCAMP) [CCAMP] makes an additional distinction between management and measurement. When the logical separation of the Control, Measurement and other Management entities is not relevant, the term Controller Plane is used for simplicity to represent them all, and the term controller refers to any device operating in that plane, whether is it a Path Computation entity or a Network Management entity (NME). The Path Computation Element (PCE) [PCE] is a core element of a controller, in charge of computing Deterministic paths to be applied in the Network Plane.

A (Northbound) Service Interface enables applications in the Application Plane to communicate with the entities in the Controller Plane.

One or more PCE(s) collaborate to implement the requests from the FME as Per-fFlow Per-Hop Behaviors installed in the relay systems for each individual flow. The PCEs place each flow along a deterministic sequence of relay systems so as to respect per-flow constraints such as security and latency, and optimize the overall result for metrics such as an abstract aggregated cost. The deterministic sequence can typically be more complex than a direct sequence and include redundancy path, with one or more packet replication and elimination points.

4.2.3. The Network Plane

The Network Plane represents the network devices and protocols as a whole, regardless of the Layer at which the network devices operate.

The network Plane comprises the Network Interface Cards (NIC) in the end systems, which are typically IP hosts, and relay systems, which are typically IP routers and switches. Network-to-Network Interfaces such as used for Traffic Engineering path reservation in [RFC3209], as well as User-to-Network Interfaces (UNI) such as provided by the Local Management Interface (LMI) between network and end systems, are all part of the Network Plane.

A Southbound (Network) Interface enables the entities in the Controller Plane to communicate with devices in the Network Plane. This interface leverages and extends TEAS to describe the physical topology and resources in the Network Plane.

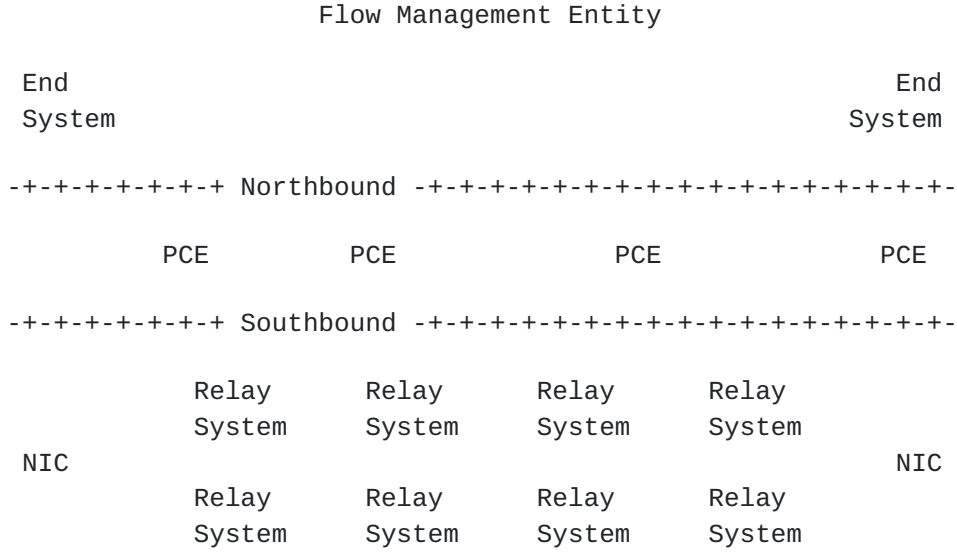

Figure 3

The relay systems (and eventually the end systems NIC) expose their capabilities and physical resources to the controller (the PCE), and update the PCE with their dynamic perception of the topology, across the Southbound Interface. In return, the PCE(s) set the per-flow paths up, providing a Flow Characterization that is more tightly coupled to the relay system Operation than a TSpec.

At the Network plane, relay systems exchange information regarding the state of the paths, between adjacent systems and eventually with the end systems, and forward packets within constraints associated to

each flow, or, when unable to do so, perform a last resort operation such as drop or declassify.

This specification focuses on the Southbound interface and the operation of the Network Plane.

4.3. DetNet flows

4.3.1. Source guarantees

DetNet flows can by synchronous or asynchronous. In synchronous DetNet flows, at least the relay systems (and possibly the end systems) are closely time synchronized, typically to better than 1 microsecond. By transmitting packets from different DetNet flows or classes of DetNet flows at different times, using repeating schedules synchronized among the relay systems, resources such as buffers and link bandwidth can be shared over the time domain among different DetNet flows. There is a tradeoff among techniques for synchronous DetNet flows between the burden of fine-grained scheduling and the benefit of reducing the required resources, especially buffer space.

In contrast, asynchronous DetNet flows are not coordinated with a fine-grained schedule, so relay and end systems must assume worst-case interference among DetNet flows contending for buffer resources. Asynchronous DetNet flows are characterized by:

- A maximum packet size;
- An observation interval; and
- A maximum number of transmissions during that observation interval.

These parameters, together with knowledge of the protocol stack used (and thus the size of the various headers added to a packet), limit the number of bit times per observation interval that the DetNet flow can occupy the physical medium.

The source promises that these limits will not be exceeded. If the source transmits less data than this limit allows, the unused resources such as link bandwidth can be made available by the system to non-DetNet packets. However, making those resources available to DetNet packets in other DetNet flows would serve no purpose. Those other DetNet flows have their own dedicated resources, on the assumption that all DetNet flows can use all of their resources over a long period of time.

Note that there is no provision in DetNet for throttling DetNet flows (reducing the transmission rate via feedback); the assumption is that a DetNet flow, to be useful, must be delivered in its entirety. That is, while any useful application is written to expect a certain number of lost packets, the real-time applications of interest to DetNet demand that the loss of data due to the network is extraordinarily infrequent.

Although DetNet strives to minimize the changes required of an application to allow it to shift from a special-purpose digital network to an Internet Protocol network, one fundamental shift in the behavior of network applications is impossible to avoid--the reservation of resources before the application starts. In the first place, a network cannot deliver finite latency and practically zero packet loss to an arbitrarily high offered load. Secondly, achieving practically zero packet loss for unthrottled (though bandwidth limited) DetNet flows means that bridges and routers have to dedicate buffer resources to specific DetNet flows or to classes of DetNet flows. The requirements of each reservation have to be translated into the parameters that control each system's queuing, shaping, and scheduling functions and delivered to the hosts, bridges, and routers.

4.3.2. Incomplete Networks

The presence in the network of relay systems that are not fully capable of offering DetNet services complicates the ability of the relay systems and/or controller to allocate resources, as extra buffering, and thus extra latency, must be allocated at points downstream from the non-DetNet relay system for a DetNet flow.

4.4. Queuing, Shaping, Scheduling, and Preemption

As described above, DetNet achieves its aims by reserving bandwidth and buffer resources at every hop along the path of the DetNet flow. The reservation itself is not sufficient, however. Implementors and users of a number of proprietary and standard real-time networks have found that standards for specific data plane techniques are required to enable these assurances to be made in a multi-vendor network. The fundamental reason is that latency variation in one system results in the need for extra buffer space in the next-hop system(s), which in turn, increases the worst-case per-hop latency.

Standard queuing and transmission selection algorithms allow a central controller to compute the latency contribution of each relay node to the end-to-end latency, to compute the amount of buffer space required in each relay system for each incremental DetNet flow, and most importantly, to translate from a flow specification to a set of

values for the managed objects that control each relay or end system. The IEEE 802 has specified (and is specifying) a set of queuing, shaping, and scheduling algorithms that enable each relay system (bridge or router), and/or a central controller, to compute these values. These algorithms include:

- A credit-based shaper [IEEE802.1Q-2014] Clause 34.
- Time-gated queues governed by a rotating time schedule, synchronized among all relay nodes [IEEE802.1Qbv].
- Synchronized double (or triple) buffers driven by synchronized time ticks. [IEEE802.1Qch].
- Pre-emption of an Ethernet packet in transmission by a packet with a more stringent latency requirement, followed by the resumption of the preempted packet [IEEE802.1Qbu], [IEEE802.3br].

While these techniques are currently embedded in Ethernet and bridging standards, we can note that they are all, except perhaps for packet preemption, equally applicable to other media than Ethernet, and to routers as well as bridges.

4.5. Coexistence with normal traffic

A DetNet network supports the dedication of a high proportion (e.g. 75%) of the network bandwidth to DetNet flows. But, no matter how much is dedicated for DetNet flows, it is a goal of DetNet to not interfere excessively with existing QoS schemes. It is also important that non-DetNet traffic not disrupt the DetNet flow, of course (see Section 4.6 and Section 7). For these reasons:

- Bandwidth (transmission opportunities) not utilized by a DetNet flow are available to non-DetNet packets (though not to other DetNet flows).
- DetNet flows can be shaped, in order to ensure that the highest-priority non-DetNet packet also is ensured a worst-case latency (at any given hop).
- When transmission opportunities for DetNet flows are scheduled in detail, then the algorithm constructing the schedule should leave sufficient opportunities for non-DetNet packets to satisfy the needs of the uses of the network.

Ideally, the net effect of the presence of DetNet flows in a network on the non-DetNet packets is primarily a reduction in the available bandwidth.

4.6. Fault Mitigation

One key to building robust real-time systems is to reduce the infinite variety of possible failures to a number that can be analyzed with reasonable confidence. DetNet aids in the process by providing filters and policers to detect DetNet packets received on the wrong interface, or at the wrong time, or in too great a volume, and to then take actions such as discarding the offending packet, shutting down the offending DetNet flow, or shutting down the offending interface.

It is also essential that filters and service remarking be employed at the network edge to prevent non-DetNet packets from being mistaken for DetNet packets, and thus impinging on the resources allocated to DetNet packets.

There exist techniques, at present and/or in various stages of standardization, that can perform these fault mitigation tasks that deliver a high probability that misbehaving systems will have zero impact on well-behaved DetNet flows, except of course, for the receiving interface(s) immediately downstream of the misbehaving device.

4.7. Protocol Stack Model

[IEEE802.1CB], Annex C, offers a description of the TSN protocol stack. While this standard is a work in progress, a consensus around the basic architecture has formed. This stack is summarized in Figure 4.

DetNet Protocol Stack

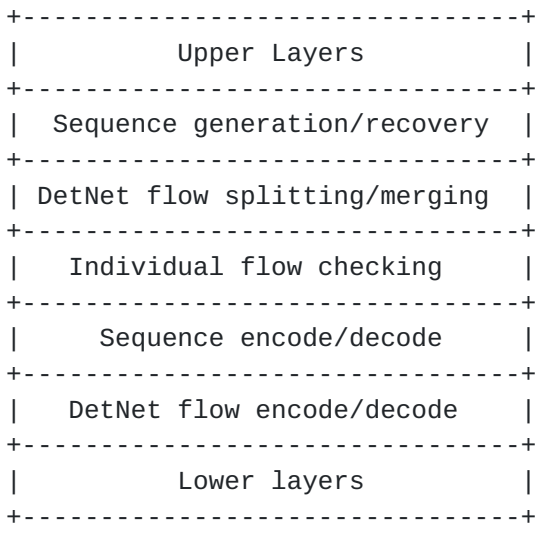

Figure 4

Not all layers are required for any given application, or even for any given network. The layers are, from top to bottom:

Sequence generation/recovery
: Supplies the sequence number for Seamless Redundancy (Section 3.3) for packets going down the stack, and discards duplicate packets coming up the stack.

DetNet flow splitting/merging
: Replicates packets going down the stack into two DetNet flows, and merges DetNet flows together for packets coming up the stack, based on the packet's DetNet flow identifier. Needed for Seamless Redundancy (Section 3.3).

Individual flow checking
: Examines packets belonging to individual flows, discards duplicate packets coming up the stack, and performs checks to detect contract violations.

Sequence encode/decode
: Encodes the sequence number into packets going down the stack, and extracts the sequence number from packets coming up the stack. This function may or may not be a null transformation of the packet, and for some protocols, is not explicitly present, being included in the DetNet flow encode/decode layer, below.

DetNet flow encode/decode

Encapsulates packets going down the stack, based on the packet's locally-significant DetNet flow identifier, in order to identify to which DetNet flow the packet belongs, and extracts a locally-significant DetNet flow identifier from packets coming up the stack. This may be a null transformation (e.g., for DetNet flows identified by IP 5-tuple) or might be an explicit encapsulation (e.g., for DetNet flows identified with an MPLS label). DetNet flow identification is the basis for Seamless Redundancy, for assigning per-flow resources (if any) to packets and for defense against misbehaving systems (Section 4.6). When DetNet flows are assigned to pinned paths, this layer can be indistinguishable from the data forwarding layer(s).

The reader is likely to notice that Figure 4 does not specify the relationship between the DetNet layers, the IP layers, and the link layers. This is intentional, because they can usefully be placed different places in the stack, and even in mulitple places, depending on where their peers are placed.

4.8. Advertising resources, capabilities and adjacencies

There are three classes of information that a central controller needs to know that can only be obtained from the end systems and/or relay systems in the network. When using a peer-to-peer control plane, some of this information may be required by a system's neighbors in the network.

- Details of the system's capabilities that are required in order to accurately allocate that system's resources, as well as other systems' resources. This includes, for example, which specific queuing and shaping algorithms are implemented (Section 4.4), the number of buffers dedicated for DetNet allocation, and the worst-case forwarding delay.

- The dynamic state of an end or relay system's DetNet resources.

- The identity of the system's neighbors, and the characteristics of the link(s) between the systems, including the length (in nanoseconds) of the link(s).

4.9. Provisioning model

4.9.1. Centralized Path Computation and Installation

A centralized routing model, such as provided with a PCE (RFC 4655 [RFC4655]), enables global and per-flow optimizations. (See

Section 4.2.) The model is attractive but a number of issues are left to be solved. In particular:

- Whether and how the path computation can be installed by 1) an end device or 2) a Network Management entity,

- And how the path is set up, either by installing state at each hop with a direct interaction between the forwarding device and the PCE, or along a path by injecting a source-routed request at one end of the path.

4.9.2. Distributed Path Setup

Whether a distributed alternative without a PCE can be valuable should be studied as well. Such an alternative could for instance inherit from the Resource ReSerVation Protocol [RFC5127] (RSVP) flows.

In a Layer-2 only environment, or as part of a layered approach to a mixed environment, IEEE 802.1 also has work, either completed or in progress. [IEEE802.1Q-2014] Clause 35 describes SRP, a peer-to-peer protocol for Layer-2 roughly analogous to RSVP. Almost complete is [IEEE802.1Qca], which defines how ISIS can provide multiple disjoint paths or distribution trees. Also in progress is [IEEE802.1Qcc], which expands the capabilities of SRP.

4.10. Scaling to larger networks

Reservations for individual DetNet flows require considerable state information in each relay system, especially when adequate fault mitigation (Section 4.6) is required. The DetNet data plane, in order to support larger numbers of DetNet flows, must support the aggregation of DetNet flows into tunnels, which themselves can be viewed by the relay systems' data planes largely as individual DetNet flows.

4.11. Connected islands vs. networks

Given that users have deployed examples of the IEEE 802.1 TSN TG standards, which provide capabilities similar to DetNet, it is obvious to ask whether the IETF DetNet effort can be limited to providing Layer-2 tunnels between islands of bridged TSN networks. While this capability is certainly useful to some applications, and must not be precluded by DetNet, tunneling alone is not a sufficient goal for the DetNet WG. As shown in the Deterministic Networking Use Cases draft [I-D.grossman-detnet-use-cases], there are already deployments of Layer-2 TSN networks that are encountering the well-

known problems of over-large broadcast domains. Routed solutions, and combinations routed/bridged solutions, are both required.

5. Compatibility with Layer-2

Standards providing similar capabilities for bridged networks (only) have been and are being generated in the IEEE 802 LAN/MAN Standards Committee. The present architecture describes an abstract model that can be applicable both at Layer-2 and Layer-3, and over links not defined by IEEE 802. It is the intention of the authors (and hopefully, as this draft progresses, of the DetNet Working Group) that IETF and IEEE 802 will coordinate their work, via the participation of common individuals, liaisons, and other means, to maximize the compatibility of their outputs.

6. Open Questions

There are a number of architectural questions that will have to be resolved before this document can be submitted for publication. Aside from the obvious fact that this present draft is subject to change, there are specific questions to which the authors wish to direct the readers' attention.

6.1. Data plane shapers and schedulers

A number of techniques have been defined and are being defined by IEEE 802 for queuing, shaping, and scheduling transmissions on EtherNet media, most of which are directly applicable to any other medium. Specific selections of supported techniques are required, because minimizing, and even eliminating, congestion losses depends strongly on the details of the per-hop behavior of sources and relay systems.

The present authors expect that, at least, the IEEE 802 mechanisms will be supported.

6.2. DetNet flow identification and sequencing

The techniques to be used for DetNet flow identification must be settled. The following paragraphs provide a snapshot of the authors' opinions at the time of writing. These authors anticipate the submission of drafts in the near future on this subject.

IEEE 802.1 TSN streams are identified by giving each stream (DetNet flow) a {VLAN identifier, destination MAC address} pair that is unique in the bridged network, and that the MAC address must be a multicast address. If a source is generating, for example, two unicast UDP flows to the same destination, one DetNet and one not,

the DetNet flow's packets must be transformed at some point to have a multicast destination MAC address, and perhaps, a different VLAN than the non-DetNet flow's packets.

A similar provision would apply to DetNet packets that are identified by MPLS labels; any bridges between the LSRs need a {VLAN identifier, destination MAC address} pair uniquely identifying the DetNet flow in the bridged network.

Provision is made in current draft of [IEEE802.1CB] to make these transformations either in a Layer-2 shim in the source end system, on the output side of a router or LSR, or in a proxy function in the first-hop bridge. It remains to be seen whether this provision is adequate and/or acceptable to the IETF DetNet WG.

There are also questions regarding the sequentialization of packets for use with Seamless Redundancy (Section 3.3). [IEEE802.1CB] defines an EtherNet tag carrying a sequence number. If MPLS Pseudowires are used with a control word containing a sequence number, the relationship and interworking between these two formats must be defined.

6.3. Flat vs. hierarchical control

Boxes that are solely routers or solely bridges are rare in today's market. In a multi-tenant data center, multiple users' virtual Layer-2/Layer-3 topologies exist simultaneously, implemented on a network whose physical topology bears only accidental resemblance to the virtual topologies.

While the forwarding topology (the bridges and routers) are an important consideration for a DetNet Flow Management Entity (Section 4.2.1), so is the purely physical topology. Ultimately, the model used by the management entities is based on boxes, queues, and links. The authors hope that the work of the TEAS WG will help to clarify exactly what model parameters need to be traded between the relay systems and the controller(s).

6.4. Peer-to-peer reservation protocol

As described in Section 4.9.2, the DetNet WG needs to decide whether to support a peer-to-peer protocol for a source and a destination to reserve resources for a DetNet stream. Assuming that enabling the involvement of the source and/or destination is desirable (see Deterministic Networking Use Cases [I-D.grossman-detnet-use-cases]), it remains to decide whether the DetNet WG will make it possible to deploy at least some DetNet capabilities in a network using only a peer-to-peer protocol, without a central controller.

7. Security Considerations

Security in the context of Deterministic Networking has an added dimension; the time of delivery of a packet can be just as important as the contents of the packet, itself. A man-in-the-middle attack, for example, can impose, and then systematically adjust, additional delays into a link, and thus disrupt or subvert a real-time application without having to crack any encryption methods employed. See [RFC7384] for an exploration of this issue in a related context.

Furthermore, in a control system where millions of dollars of equipment, or even human lives, can be lost if the DetNet QoS is not delivered, one must consider not only simple equipment failures, where the box or wire instantly becomes perfectly silent, but bizarre errors such as can be caused by software failures. Because there is essential no limit to the kinds of failures that can occur, protecting against realistic equipment failures is indistinguishable, in most cases, from protecting against malicious behavior, whether accidental or intentional. See also Section 4.6.

Security must cover:

- the protection of the signaling protocol
- the authentication and authorization of the controlling systems
- the identification and shaping of the DetNet flows

8. IANA Considerations

This document does not require an action from IANA.

9. Acknowledgements

The authors wish to thank Jouni Korhonen, Erik Nordmark, George Swallow, Rudy Klecka, Anca Zamfir, David Black, Thomas Watteyne, Shitanshu Shah, Craig Gunther, Rodney Cummings, Wilfried Steiner, Marcel Kiessling, Karl Weber, Ethan Grossman, Pat Thaler, and Lou Berger for their various contribution with this work.

10. Access to IEEE 802.1 documents

To access password protected IEEE 802.1 drafts, see the IETF IEEE 802.1 information page at https://www.ietf.org/proceedings/52/slides/bridge-0/tsld003.htm.

11. Informative References

[AVnu] http://www.avnu.org/, "The AVnu Alliance tests and certifies devices for interoperability, providing a simple and reliable networking solution for AV network implementation based on the Audio Video Bridging (AVB) standards.".

[CCAMP] IETF, "Common Control and Measurement Plane", <https://datatracker.ietf.org/doc/charter-ietf-ccamp/>.

[HART] www.hartcomm.org, "Highway Addressable Remote Transducer, a group of specifications for industrial process and control devices administered by the HART Foundation".

[HSR-PRP] IEC, "High availability seamless redundancy (HSR) is a further development of the PRP approach, although HSR functions primarily as a protocol for creating media redundancy while PRP, as described in the previous section, creates network redundancy. PRP and HSR are both described in the IEC 62439 3 standard.", <http://webstore.iec.ch/webstore/webstore.nsf/artnum/046615!opendocument>.

[I-D.finn-detnet-problem-statement] Finn, N. and P. Thubert, "Deterministic Networking Problem Statement", draft-finn-detnet-problem-statement-04 (work in progress), October 2015.

[I-D.grossman-detnet-use-cases] Grossman, E., Gunther, C., Thubert, P., Wetterwald, P., Raymond, J., Korhonen, J., Kaneko, Y., Das, S., and Y. Zha, "Deterministic Networking Use Cases", draft-grossman-detnet-use-cases-01 (work in progress), November 2015.

[I-D.ietf-6tisch-architecture] Thubert, P., "An Architecture for IPv6 over the TSCH mode of IEEE 802.15.4", draft-ietf-6tisch-architecture-09 (work in progress), November 2015.

[I-D.ietf-6tisch-tsch] Watteyne, T., Palattella, M., and L. Grieco, "Using IEEE802.15.4e TSCH in an IoT context: Overview, Problem Statement and Goals", draft-ietf-6tisch-tsch-06 (work in progress), March 2015.

[I-D.ietf-roll-rpl-industrial-applicability]
Phinney, T., Thubert, P., and R. Assimiti, "RPL applicability in industrial networks", draft-ietf-roll-rpl-industrial-applicability-02 (work in progress), October 2013.

[I-D.svshah-tsvwg-deterministic-forwarding]
Shah, S. and P. Thubert, "Deterministic Forwarding PHB", draft-svshah-tsvwg-deterministic-forwarding-04 (work in progress), August 2015.

[IEEE802.1AS-2011]
IEEE, "Timing and Synchronizations (IEEE 802.1AS-2011)", 2011, <http://standards.ieee.org/getIEEE802/download/802.1AS-2011.pdf>.

[IEEE802.1BA-2011]
IEEE, "AVB Systems (IEEE 802.1BA-2011)", 2011, <http://standards.ieee.org/getIEEE802/download/802.1BA-2011.pdf>.

[IEEE802.1CB]
IEEE, "Seamless Redundancy (IEEE Draft P802.1CB)", 2016, <http://www.ieee802.org/1/files/private/cb-drafts/>.

[IEEE802.1Q-2014]
IEEE, "MAC Bridges and VLANs (IEEE 802.1Q-2014", 2014, <http://standards.ieee.org/getIEEE802/download/802.1Q-2014.pdf>.

[IEEE802.1Qbu]
IEEE, "Frame Preemption", 2016, <http://www.ieee802.org/1/files/private/bu-drafts/>.

[IEEE802.1Qbv]
IEEE, "Enhancements for Scheduled Traffic", 2016, <http://www.ieee802.org/1/files/private/bv-drafts/>.

[IEEE802.1Qca]
IEEE, "Path Control and Reservation", 2015, <http://www.ieee802.org/1/files/private/ca-drafts/>.

[IEEE802.1Qcc]
IEEE, "Stream Reservation Protocol (SRP) Enhancements and Performance Improvements", 2016, <http://www.ieee802.org/1/files/private/cc-drafts/>.

[IEEE802.1Qch]
IEEE, "Cyclic Queuing and Forwarding", 2016, <http://www.ieee802.org/1/files/private/ch-drafts/>.

[IEEE802.1TSNTG]
IEEE Standards Association, "IEEE 802.1 Time-Sensitive Networks Task Group", 2013, <http://www.IEEE802.org/1/pages/avbridges.html>.

[IEEE802.3br]
IEEE, "Interspersed Express Traffic", 2016, <http://www.ieee802.org/3/br/>.

[IEEE802154]
IEEE standard for Information Technology, "IEEE std. 802.15.4, Part. 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks", June 2011.

[IEEE802154e]
IEEE standard for Information Technology, "IEEE std. 802.15.4e, Part. 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs) Amendment 1: MAC sublayer", April 2012.

[ISA100.11a]
ISA/IEC, "ISA100.11a, Wireless Systems for Automation, also IEC 62734", 2011, < http://www.isa100wci.org/en-US/Documents/PDF/3405-ISA100-WirelessSystems-Future-broch-WEB-ETSI.aspx>.

[ISA95] ANSI/ISA, "Enterprise-Control System Integration Part 1: Models and Terminology", 2000, <https://www.isa.org/isa95/>.

[ODVA] http://www.odva.org/, "The organization that supports network technologies built on the Common Industrial Protocol (CIP) including EtherNet/IP.".

[PCE] IETF, "Path Computation Element", <https://datatracker.ietf.org/doc/charter-ietf-pce/>.

[Profinet]
http://us.profinet.com/technology/profinet/, "PROFINET is a standard for industrial networking in automation.", <http://us.profinet.com/technology/profinet/>.

[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, <http://www.rfc-editor.org/info/rfc2205>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <http://www.rfc-editor.org/info/rfc3209>.

[RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <http://www.rfc-editor.org/info/rfc4655>.

[RFC5127] Chan, K., Babiarz, J., and F. Baker, "Aggregation of Diffserv Service Classes", RFC 5127, DOI 10.17487/RFC5127, February 2008, <http://www.rfc-editor.org/info/rfc5127>.

[RFC5673] Pister, K., Ed., Thubert, P., Ed., Dwars, S., and T. Phinney, "Industrial Routing Requirements in Low-Power and Lossy Networks", RFC 5673, DOI 10.17487/RFC5673, October 2009, <http://www.rfc-editor.org/info/rfc5673>.

[RFC7384] Mizrahi, T., "Security Requirements of Time Protocols in Packet Switched Networks", RFC 7384, DOI 10.17487/RFC7384, October 2014, <http://www.rfc-editor.org/info/rfc7384>.

[RFC7426] Haleplidis, E., Ed., Pentikousis, K., Ed., Denazis, S., Hadi Salim, J., Meyer, D., and O. Koufopavlou, "Software-Defined Networking (SDN): Layers and Architecture Terminology", RFC 7426, DOI 10.17487/RFC7426, January 2015, <http://www.rfc-editor.org/info/rfc7426>.

[TEAS] IETF, "Traffic Engineering Architecture and Signaling", <https://datatracker.ietf.org/doc/charter-ietf-teas/>.

[WirelessHART] www.hartcomm.org, "Industrial Communication Networks - Wireless Communication Network and Communication Profiles - WirelessHART - IEC 62591", 2010.

Authors' Addresses

Norman Finn
Cisco Systems
170 W Tasman Dr.
San Jose, California 95134
USA

Phone: +1 408 526 4495
Email: nfinn@cisco.com

Pascal Thubert
Cisco Systems
Village d'Entreprises Green Side
400, Avenue de Roumanille
Batiment T3
Biot - Sophia Antipolis 06410
FRANCE

Phone: +33 4 97 23 26 34
Email: pthubert@cisco.com

Michael Johas Teener
Broadcom Corp.
3151 Zanker Rd.
San Jose, California 95134
USA

Phone: +1 831 824 4228
Email: MikeJT@broadcom.com