ATIS-I-0000046

Emerging Opportunities for Leveraging Network Intelligence

October 2014

1 Executive Summary

Service provider networks possess a wealth of information (network intelligence) relative to subscribers, applications and network state. Attempts to leverage this intelligence have, in the past, fallen short. The potential gain associated with network intelligence is significant. Additionally, new network technologies such as NFV and SDN, provide new possibilities for network and application optimization.

This focus group report has taken a “use case” approach in analyzing how network intelligence can be better leveraged in the network. Eleven different use cases have been analyzed. For each use case, both the network data inputs as well as potential outputs/network actions were considered to better leverage network intelligence.

Standards gaps were identified in three areas:

- Application specific data collection.
- Network data collection.
- Network actions or control points that can be used as a result of data analysis.

Additionally, three common themes emerged as a result of the analysis. First, the correlation of input data can greatly improve the range and effectiveness of subsequent actions. For example, a node, link or device may report congestion. If the congestion event can be correlated to identify specific subscribers participating in the congestion and if the subscriber contribution to congestion can be further correlated based on what application is being used, there is a much richer set of potential mitigations that can better optimize network resources as well as create new monetization opportunities.

The second theme is that the emerging NFV and SDN work creates new opportunities to utilize network information by enabling the implementation of automated network mitigations to “automatically” adjust network configuration and parameters to deal with changing traffic patterns and call models.

Finally, the Focus Group concluded that a common data exposure framework could increase the availability and usability of network data as collected by various analytics systems. That is, it would be useful if network elements could natively expose the right data in a common and consistent manner to allow analytics systems to capture the data with standard API calls within the network element. The focus group also realizes that such an undertaking may be a significant industry effort, however; this may be an opportune time to address this as network elements are adapted for more programmable control. In any case, consistent and timely access to data is essential to analytics analysis.

2 Introduction

Service provider networks possess a wealth of information (network intelligence) relative to subscribers, applications and network state. As service providers look to increase the value delivered to their subscribers and differentiate their fixed and mobile broadband services, they need to optimize the value of their network resources as well as leverage the massive amount of rich network data they possess.

Unfortunately, attempts to leverage this intelligence have, in the past, fallen short. A great deal of network intelligence is embedded in network elements with no standard or cost-effective mechanism to extract the information. Even when this information can be extracted, there often is no means to properly correlate extracted data across the network to create useful abstractions that can be used to both optimize network utilization as well as increase overall subscriber Quality of Experience (QoE).

Yet, the potential gain associated with network intelligence is significant. Additionally, new network technologies such as NFV and SDN, provide new possibilities for network and application optimization.

Network Intelligence exists across the network and tends to fall into three main categories:

- Subscriber data or Identity-based information may be used to make services and applications personalized and more attractive for the end-user. Personalization is a key component in

delivering enhanced customer experiences leading to improved customer satisfaction while managing network utilization.

- Application information allows the network to implement application specific optimizations and mitigations to increase overall subscriber QoE. Different applications have different delivery needs, either conversational real time, near real time or background in nature. By understanding these needs, a network can better provide reasonable network management practices in the face of congestion to optimize subscriber QoE while efficiently managing network resources.
- Network state information may include information such as congestion status of key choke points in the network including access/RAN aspects as well as transport link level aspects. Additionally, traffic statistics as well as the security posture of the user device and associated network elements can provide a complete end-to-end view of the network environment.

The deployment of SDN and NFV technologies provides new network control mechanisms that can leverage network intelligence to advantage. Previously, even if it was possible to gather real-time network information, the network could not easily take immediate action. With real-time analytics, an SDN controller can have a complete picture of the network. The controller already has global control of network state. This combination enables opportunities to dynamically optimize network utilization and application delivery.

With NFV, it is now possible to more dynamically size the network to quickly accommodate changes in call models and traffic patterns. Network operators can also better respond to security issues by instantiating security appliances at specific points in the network to address specific situations.

In the past, ATIS has studied various network optimization and big data analytics topics. In September 2011, ATIS published a report titled, “Network Optimization Focus Group (NetOp-FG) Assessment and Recommendations”. In this report, ATIS considered 7 different network optimization use cases and identified specific standards gaps to be worked. In October of 2013, ATIS published a report titled, “ATIS-I-0000043 - ATIS Big Data Analytics Focus Group: BDA Data Value Chain Reference Model & Use Cases”. This document provided common terminology, typical use cases, best practices, most appropriate technologies, and emerging developments as they apply to service provider network analytics.

This focus group report takes a “use case” approach in analyzing how network intelligence can be better leveraged in the network. Section 6 documents a wide range of use cases. For each use case, each element/function in the network applicable to the use case to identify potential information elements that may be leveragable was systematically reviewed. The report specifically looks closely at how information can be correlated across the network to provide more useful data abstractions. The report also looks at ways in which the data can be used to create actions in the network to effect the optimization targeted by the use case.

The analysis includes a view of potential standards gaps. Standards opportunities might exist in the collection of real-time information from network elements up to a control/presentation layer (data definitions as well as an interface between the layers). Providing this information in a standardized format will create opportunities to use analytics to extract additional value from this data.

3 DEFINITIONS

For a list of common communications terms and definitions, please visit the ATIS Telecom Glossary, which is located at < http://www.atis.org/glossary>.

ABR: Adaptive Bit Rate (typically applied to video)

ACL: Access Control List

API: Application Programming Interface

ARP: Allocation and Retention Priority (3GPP context)

BSSID: Basic Service Set Identification – typically the MAC address of a wireless access point

CAC: Call Admission Control

CDN: Content Delivery Network

DC: Data Center

DDoS: Distributed Denial-of-Service

DHCP: Dynamic Host Configuration Protocol

DNS: Domain Name Server

DPI: Deep Packet Inspection

DSCP: Differentiated Services Code Point

eMPS: Enhanced Multimedia Priority Service

eNB: evolved NodeB – an LTE basestation

EPC: Evolved Packet Core

ETSI: European Telecommunications Standards Institute

GBR: Guaranteed Bit Rate

GETS: Government Emergency Telecommunications Service

GiLAN: Refers to the Internet side of a mobility access gateway IP anchor point

HSS: Home Subscriber Server

HTTP 2.0: Planned next version of HTTP - Hypertext Transfer Protocol

IE: Information Element

IPS: Intrusion Protection System

KPI: Key Performance Indicator

M2M: Machine to Machine type communication

MAC address: Media Access Control Address

MANO: NFV Management and Orchestration

MPLS: Multi-Protocol Label Switching

NE: Network Element

NFV: Network Functions Virtualization

NGN: Next Generation Network

OAM: Operations, Administration and Management

OCS: Online Charging System

OFCS: Off Line Charging System

OSS: Operations Support System

OTT: Over-The-Top

P-GW: PDN (Public Data Network) Gateway

PCEF: Policy and Charging Enforcement Function

PCRF: Policy and Charging Rules Function

PS: Public Safety

PSAP: Public Safety Answering Point

QoE: Quality of Experience

QoS: Quality of Service

RAN: Radio Access Network

SBC: Session Border Controller

SDN: Software Defined Network

SDO: Standards Development Organization

SIP: Session Initiation Protocol

SLA: Service Level Agreement

SON: Self Organizing Networks

SP: Service Provider

SPDY: Open networking protocol developed primarily at Google for transporting web content.

SPR: Subscriber Policy Repository

STB: Set Top Box

TCP: Transmission Control Protocol

TSP: Telecommunications Service Priority

UDR: User Data Repository

UE: User Equipment

VoIP: Voice over IP

VM: Virtual Machine

VNF: Virtual Network Function

VOD: Video On Demand

VPN: Virtual Private Network

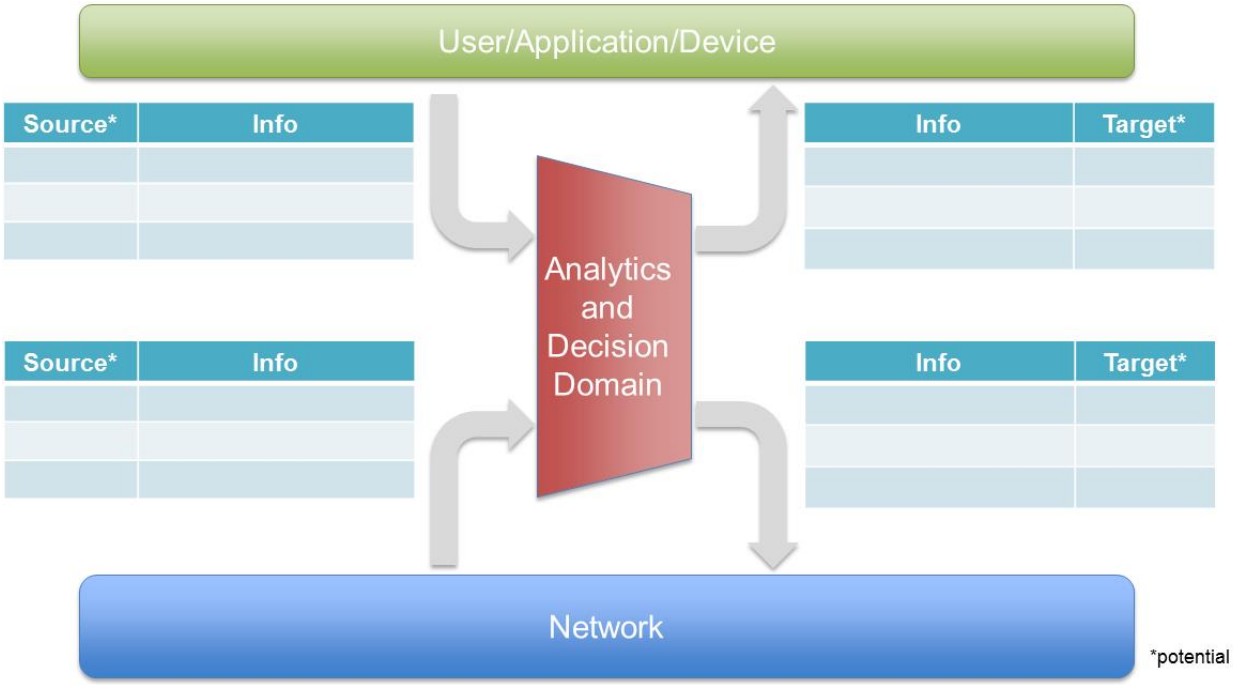
WAN: Wide Area Network

4 Analysis Model & Methodology

4.1 Framework for Analysis

A “use case” approach has been employed in analyzing how network intelligence can be better leveraged in the network. Examples of beneficial uses of network intelligence include:

- User Experience Optimization – rapid QoE impairment detection & mitigation and congestion management & mitigation mechanisms;
- Network Optimization – optimization of network asset utilization, demand time shifting to off-peak, and other operations efficiency examples;
- Monetization Optimization - proactive feature offerings such as bandwidth upgrades, network resource exposure & consumption (e.g., via APIs) to 3rd application providers, and dynamic pricing/charging offers; and
- Network Integrity - proactive maintenance, rapid outage restoration, graceful disaster recovery and outage reporting mechanisms.



A common template for use case creation is shown in the above figure. In all cases, it is assumed that:

- Information is collected from either the network or the user/application/device and aggregated by one or more "Analytics" functions.
- The resulting analysis applies use case specific (application specific) logic to create an output action. Note that this action can be a change in network behavior, a change in application behavior, or a data report that can be used offline by operations and/or marketing organizations within the network provider domain.

After documenting a set of applicable use cases, the aggregate set of common input information elements and output actions was analyzed relative to:

- Commonality of both input information elements and output actions across all use cases to assess the most common (and therefore most useful) elements and actions;
- Data correlation requirements to meet the use case needs; and
- Standardization gaps.

4.2 Identification & Evaluation of Valuable Use Cases

The following use cases are further analyzed in Section 6:

1. Network-wide Application Detection and Usage Support.
2. Prioritization of Traffic.
3. Personalized Broadband.
4. Public Safety Spectrum Sharing.
5. Enhanced Fault Resolution.
6. Outage Alerting, Avoidance & Reporting.
7. Network Wide Intrusion Detection.
8. NFV Automated Network Growth/Degrowth.
9. Inter-DC Congestion Mitigation.
10. RAN-Aware Time-Shifted Content Delivery.
11. Dynamically inspect traffic and predict performance of the network.

5 New Opportunities & Barriers to Leverage Network Intelligence

5.1 New Network Technologies & Architectures

Since the initial work in ATIS on network optimization in 2011, a number of new technologies have gained deployment traction within service provider networks. These technologies include big data analytics, SDN and NFV, and the introduction of SPDY and HTTP 2.0 protocols to web based traffic. These technologies all impact the network service provider's ability to collect and leverage network intelligence and to provide reasonable network management within the network.

5.2 Big Data Analytics

Three factors have combined to make Big Data Analytics a reality. The price of storage has fallen precipitously, processing power has increased enormously and the amount of data being created by networks, applications, and devices has grown to an exabyte every day. Capturing and analyzing this data has become a very robust business in a very short time: $12 billion in 2013 alone. Big data analytics has become a mainstream goal in the communications industry. Yet, the sharing and application/usage of analytics insight with multiple parties is in its nascent stages. Service providers are at different maturity levels of applying big data analytics.

There are three key data-related issues that need to be addressed and defined within the domain of big data analytics:

- What is the information that is being gathered?
- Where is it coming from?
- How is the information allowed to be used?

In this report, the focus is on the information gathering aspects of big data analytics to ensure that useful data is easily available in a standard format with the ability to correlate different data sets to extract even more insight about the network and the applications in use.

5.3 SDN & NFV

SDN and NFV are different but strongly related technologies. SDN separates control (the SDN Controller) from the data plane aspects in the network. By separating and centralizing control, routing/forwarding decisions can now be programmatic and take better advantage of network data and analytics. This separation of control and data provides a new and powerful mechanism to leverage network intelligence.

NFV is fundamentally about the separation of software from the underlying hardware. This is done by creating a virtualization infrastructure that allows software to run as a virtual machine on top of standard compute, storage, and networking assets. This separation of software from hardware allows an NFV management and orchestration function to instantiate functions in the network where they need to be while allowing the functions to grow and shrink dynamically based on network conditions. This flexibility in adapting network configuration and operating points provides another opportunity and mechanism to better leverage network intelligence.

Both NFV and SDN are of great interest to the ATIS member companies. In October 2013, ATIS produced a focus group report entitled, "Operational Opportunities and Challenges of SDN/NFV Programmable Infrastructure." ATIS continues to consider the impacts of NFV/SDN in service provider networks with additional focus groups and forums.

5.4 SPDY & SPDY Proxies

Web content is becoming more complex both in size and in the number of separate connections/requests. Web pages are dominated by images and scripts, both of which are easily compressible, but are not

compressed by current implementations. Content is increasingly being delivered to mobile devices over wireless networks which are constrained in terms of bandwidth and latency. Unfortunately, protocols like HTTP (application level request/response) and TCP (reliable transport) were not designed to handle today's larger, composite web pages, and performance, especially page load times, suffers as a result.

These factors provide motivation to speed up the web (performance optimization), especially to mobile devices. There is broad industry consensus that web page delivery optimization will be necessary to continue to deliver a quality user experience. Building on early deployments of such optimizations, the Internet Engineering Task Force (IETF) has embraced the SPDY protocol as a starting point for standardization of HTTP/2.0 (expected Nov 2014).

SPDY, described as "an experimental protocol for a faster web", is a new protocol being introduced by Google to reduce web page load times. Deployment is increasing, but as of today many web servers still do not support SPDY. To accelerate adoption of the protocol, SPDY proxy services have been deployed to offer some of the benefits of SPDY even when downloading content from web servers that have not yet implemented the SPDY protocol. However, by bundling multiple requests and encrypting all web traffic for a user inside a single connection, the SPDY proxy creates an opaque tunnel, hiding the true source of the content and breaking reasonable network management, content distribution, and services offered by the network operators. This would be equally true for any proxy service that bundled all traffic from a given user inside an opaque tunnel, irrespective of the protocol used.

The increasing use of Google's Data Compression Proxy and similar non-standard/proprietary proxy deployments for web traffic impacts the operator's ability to leverage network intelligence, particularly with respect to reasonable network management practices such as traffic optimization (compression proxies likely add one or more hop), malware detection, content caching, DNS resolution and content filtering to enforce parental control. While some of the use cases analyzed in this study could be impacted by these trends, the analysis of ways of addressing or mitigating these issues are the subject of parallel ATIS-related activities, and are out of scope for this paper.

This topic is further explored in the ATIS white paper from April 2014 titled "An Analysis of the SPDY Protocol and the SPDY Proxy". Additionally, ATIS has sponsored a new forum called the "Open Web Alliance" (OWA) to develop requirements for an open service optimization proxy to meet the service needs of all stakeholders in the web ecosystem while supporting the goals of encryption and privacy. The initiative brings together all organizations with a stake in making integrated web service richer, faster and more secure to address the challenges SPDY brings. Membership is open to all who desire a more equitable and workable web.

6 Use Case Analysis

6.1 Use Case 1: Network-wide Application Detection & Usage Support

6.1.1 Description & Business Value

The goal of the Network-Wide Application Detection and Usage Support use case is to collect application specific performance metrics, combine these with network metrics (e.g., congestion), and process these metrics for the purpose of improving overall application/network performance. This is done by passing relevant application information to both:

- Application specific network functions (e.g., media optimization functions) to enable network based optimization for that application and/or;
- The application function itself to enable application specific optimizations to improve application performance.

Consider for example a service provider operated video service. Today, the content industry does not have a universal, standardized tool-kit approach to solving the diverse needs of users – particularly mobile users. Specifically, video formats, delivery, and optimization methods are fragmented and often

highly proprietary. As such, video QoE in many networks scenarios (specifically mobile networks) is problematic. Many video application providers have designed 'hooks' into their native players to try to address this problem in silos.

However, a more comprehensive and standardized approach could be implemented through the collection of Video QoE metrics from devices along with associated network data. Essentially, device and network analytics could be gathered and used for the purpose of being "Mobile Network Aware" and "Content Aware". In this context, video encoding / transcoding, translating, and optimization could be done based on this intelligence.

From an overall business value perspective, application awareness provides the opportunity for the network provider to serve the subscriber with higher QoE. In the case where network assets are shared, it may be possible to increase overall QoE of the affected user population (not just users of a specific application).

6.1.2 Variations

In addition to the video optimization case sited above, there are many different possible uses for this type of application intelligence. For example:

- Generate reports on multiple levels (application usage, application usage per subscriber group, application ranking.
- Correlation of ranking with revenue generation.
- Determine broader network/subscriber policies on usage reports and service provider business logic.

6.1.3 Input Data

The following device/client data would be useful for the video example cited above:

- Used Bandwidth.
- ABR Rate Selection.
- Audio, Video.
- Stalls.
- Total number of TCP Connections.
- Packet Size (ABR Video).
- Packet Arrival Time.

Network data for this example could include:

- Location (e.g., for a mobile, the cell / sector or BSSID for Wi-Fi).
- Congestion state at that location and/or along the forwarding path.
- Subscriber identity and associated subscription profile.

6.1.4 Output Action & Target

Application based actions could (for example) be facilitated by presenting the results to a network video analytics gateway. This gateway could have APIs that would allow application servers for determine specific performance parameters for the sessions of interest.

Network based actions could be taken by network specific media optimization devices and/or CDNs which could modify delivery of the video to better suit network conditions.

6.1.5 Data Correlation Needs

To properly implement this use case, certain aspects of the input data need to be correlated. For example:

- Application Type (video for the example sited) should be correlated with application usage statistics (e.g., bandwidth and volume). This can be more easily accomplished when the application usage data is collected by the application itself. In the video example cited above, usage data was collected by the device client who was intimately aware of the application.
- The application type and usage information should be correlated with the Subscriber ID of the user. This allows subscription profile information to be leveraged in the decision making process.
- The application type and usage information should also be correlated to the user's location. This allows the application or network to look for broad location based trends and accommodate these in the decision making process.
- Finally the location information should be correlated with network congestion state for that location. Knowing whether a specific location is or is not in some level of congestion provides for better decision making when choosing initial and maximum video rates in ABR applications.

6.1.6 Gaps

The key gap associated with this use case is indicated by the data correlation needs of the previous section. Today, it is often difficult to correlate diverse data items that are often collected by systems that do not have the ability to associate the data with other critical information in the network.

6.2 Use Case 2: Prioritization of Traffic

6.2.1 Description & Business Value

The goal of the Prioritization of Traffic use case it to allocate network resources under network congestion conditions in such a way as to ensure:

- Maximum utilization of the network;
- User QoE is maintained at acceptable levels;
- SLAs are met;
- Revenue is maximized; and
- Regulatory commitments are met.

Consider for example a Service Provider offering 3rd party/OTT services to their users. The SP and the OTT operator have SLAs with regards to the sponsors, QoS, charging, revenue sharing, maximum number of simultaneous connections etc. While the SP must ensure that SLAs with all OTT providers are met a solution for resource allocation is needed that guarantees maximum utilization of the network and maximum revenue. Such a solution must assign a relative priority level to each OTT provider that is reflected in the priority and QoS derived for the OTT connection to a UE.
Another example would be a SP offering eMPS (NGN GETS in North America) and is obligated to handle with priority voice, video, and data connections to/from eMPS users. eMPS service can be invoked at connection/session set-up time or in mid-session.

6.2.2 Variations

In addition to the examples given above case sited above, there are many different possible uses for this type of prioritization of traffic. For example:

- Prioritization of SP vs. OTT applications.
- Subscriber QoS Profile (Gold, Silver, Bronze) and usage levels (e.g., heavy user).
- Roaming vs. non-roaming traffic.
- GBR vs. Non-GBR traffic.

6.2.3 Input Data

6.2.3.1 Policy Management Interface

- OTT/3rd Part Application domain.
  - OTT priority.
  - Connection priority.
- SPR/UDR.
  - OTT provider profile.
  - Allowed applications and priority.
  - Current usage measurement.

6.2.3.2 Network Utilization & Performance levels

- Location (e.g., for a mobile, the cell / sector or BSSID for Wi-Fi).
- Congestion state at that location and/or along the forwarding path.
- eNB ARP Handling KPIs.

  The ARP assigned to the connection ultimately determines admission control, pre-emption, and successful handover of a bearer if the target eNB is under overload/congested.

  - Connections pre-empted (where allowed per regulatory requirements) to admit a new connection with higher priority level.
  - Current Number of non-pre-emptible connections.
  - Connection downgraded due to congestion in the previous time interval in order to maintain the QoS of connections with higher ARP.
  - Bearers not accepted for hand over by the target eNB.
- SON – Measurements per priority level (of the ARP component) per Cell/Location.

6.2.4 Output Action & Target

Prioritization of traffic could be facilitated if the information elements (IEs) listed in the previous section are available at the PCRF.

The PCRF takes into account priority IEs it receives from the application domain and the SPR in determining the ARP of the connection under normal network load conditions.

Under overload conditions the PCRF takes into account the priority level measurements derived by SON in:

1. Determining the ARP of the connection.
2. Downgrading the ARP of existing connections.
3. Throttling.
4. Content Optimization.

6.2.5 Data Correlation Needs

To properly implement this use case, certain aspects of the input data need to be correlated. For example:

In order to decide whether to admit a new user connection and, if admitted, the ARP of that connection, PCRF needs to correlate:

1. Priority of the OTT and the priority level of the requested connection and with the OTT's QoS profile in the SPR and allowed applications.
2. Usage measurements per OTT and usage measurements per OTT applications.
3. Congestion level of the location/cell and user's current location.
4. Measurements per priority level (of the ARP component) per Cell/Location.

6.2.6 Gaps

PCC interfaces:

- o New IEs over the Rx interface.
- o SPR/UDR profile for OTT/3rd party application provided.

UPCON solution supports sending to the PCRF:

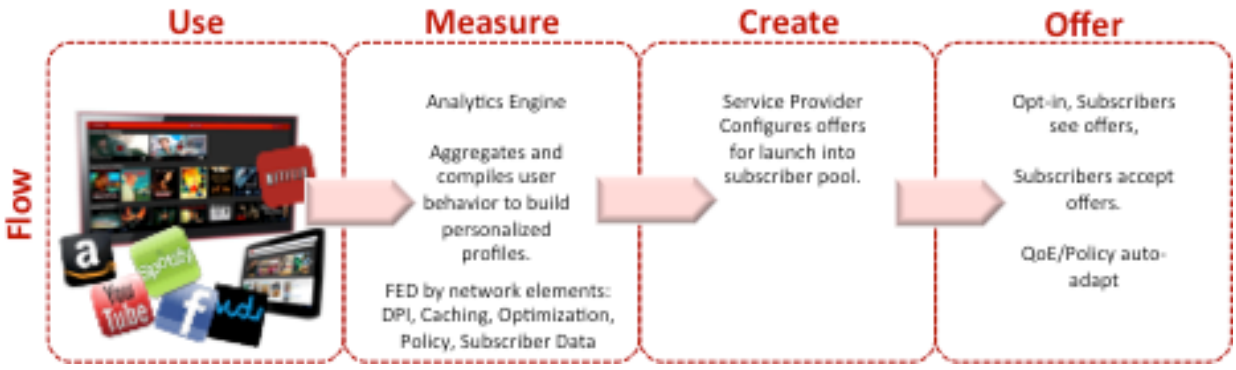
- o Cell Congestion Level.
- o ARP KPIs per Cell.

6.3 Use Case 3: Personalized Broadband

6.3.1 Description & Business Value

Network service providers are challenged today with maximizing data service plan revenues and minimizing subscriber churn. Several existing solutions have sought to shore up Internet data service plan revenues and better align revenues with data usage. However, these plans may not be well aligned to the subscriber's historical and planned access to Internet Protocol (IP)-based content and services, and cannot adapt to the subscriber's changing usage patterns.

Personalized Broadband provides the ability for a service provider to offer customized broadband service plans that meet the needs of individual subscribers thereby enabling incremental improvements to data-service subscription revenue and costs. The Personalized Broadband solution flow is described below:



6.3.2 Variations

There are many different possible applications for personalized broadband. Some examples include:

- Capped and tiered data (quota) plans.
- Enhanced Streaming QoE.
- Enhanced Data QoE.
- Off Peak “happy hour”.
- Gaming (delay sensitive).
- Service Specific Offers.
- Zero Rated OTT Services.
- Casual Usage / Loyalty.

6.3.3 Input Data

Subscriber profile and network usage data for a subscriber are used to generate a personalized data service offer.

Subscriber data includes current subscription options, user authentication data, and user device configuration data. Sources of subscriber data may be include one or more network elements including Authentication, Authorization, and Accounting (AAA) server, Home Subscriber Server (HSS), Subscriber Policy Repository (SPR), DHCP server, proprietary policy repository associated with a DPI inline engine, or a User Profile Server Function (UPSF).

Network usage data includes user behavior, application visibility, and traffic patterns on the network at certain times of day, week month or year. Inline engines may provide the source of this data either as standalone DPI network element or integrated into an access gateway, such as a Broadband Remote Access Server (BRAS), Broadband Network Gateway (BNG), Cable Modem Termination System (CMTS), Packet Data Network (PDN) Gateway, Gateway GPRS Serving Node (GGSN), Home Agent, Packet Data Serving Node (PDSN), or the like.

6.3.4 Output Action &Target

Notifications toward users alerting them on bespoke service offering or redirecting their web browsing session to a view an operator promotion.

Personalized data service offers that lead to changes in a subscriber's service configuration are propagated to the network upon acceptance by the subscriber. The following actions may be triggered:

- Update of subscription data in subscriber data repositories.
- Policy Server distributes appropriate policy updates to applicable network elements for enforcement. Examples include:
    - PCRF may send wireline QoS parameters to Router and wireless QoS parameters to P-GW.
    - PCRF may send traffic shaping/gating/marking/redirection/ traffic optimization parameters to DPI engine.
- Policy Server updates of charging/rating/billing systems to apply correct charging for subscriber.

6.3.5 Data Correlation Needs

To implement Personalized Broadband the profile and network usage data must be correlated:

- Historical network usage of the subscriber.
- Statistical network usage data of the subscriber.
- Current subscription information of the subscriber.
- Congestion status during a given traffic period.
- Degree of mobility of the user.

6.3.6 Gaps

This solution breaks down into:

- Information gathering.
- Information normalization.
- Information processing.
- User interaction.
- Network provisioning.
- Enforcement.

Each stage can be achieved to some degree with standardized or internal mechanisms. The combination of all these aspects to deliver the service described will most likely require some degree of bespoke integration and potentially enhancement of standard mechanisms or proprietary extensions.

6.4 Use Case 4: Public Safety Spectrum Sharing

6.4.1 Description & Business Value

The efficient utilization of wireless spectrum for all types of communications purposes is of utmost importance to a country's economic opportunity growth, and to the regulatory bodies (e.g., the FCC in the US) charged with overseeing spectrum usage. Traditionally, governmental users such as the military or public safety entities have been allocated spectrum dedicated for their exclusive use. As all types of users (including public safety users) consume ever more sophisticated broadband services, with higher definition and increasingly video content, such inefficient balkanization of spectrum is inconsistent with meeting the superset of these new burgeoning demands. It also leads to expensive network buildouts for special networks, hard to prove in for usage that may have low duty cycles in time and space. If new and/or existing bands of spectrum that would otherwise be allocated/dedicated to special use, can instead be efficiently shared across governmental and commercial use, costs can be contained, coverage can be maximized, and all user groups stand to gain.

At the same time, new technologies for building and automating networks – particularly SDN and NFV – enable unprecedented levels of flexible network programmability which can be leveraged for highly dynamic sharing of spectrum between different entities and/or various classes of users. This includes flexible ways of establishing monitoring points and collecting information to increase visibility into what is happening within the network, to specific users' experience, and even to the broader environment. It also includes flexible controls to dynamically adjust resources in the network, automatically or semi-automatically reacting to the current situation.

One example of spectrum sharing is the FCC rulemakings regarding sharing of spectrum in the 3.5 GHz band currently dedicated to government use (military radars and other uses). Another example is the establishment of larger blocks of spectrum for public safety broadband, allowing emergency personal unblocked access to rich communications resources wherever and whenever needed. As a prime example of an opportunity to efficiently share spectrum, geographic coverage and network costs – the Focus Group will further develop this specific example for purposes of this use case analysis. Many of the concepts outlined in this specific sharing use case can be applied to other spectrum sharing opportunities.

In this public safety (PS) spectrum sharing use case, PS entities would leverage the high capacity and broad coverage of commercial LTE wireless network(s) enhanced by access to the additional spectrum made available for PS shared use. Under normal circumstances, PS users can utilize this enhanced network, taking advantage (along with all users) of the combined resources of the PS shared as well as normal licensed spectrum. Under emergency or disaster circumstances, these spectrum and associated network resources can be shifted to the exclusive use of PS users for the affected geography.

6.4.2 Variations

This use case can take many different forms and levels of sophistication, for both the input or event detection side, as well as in the form of action taken in response to an emergency situation. Some examples of each are outlined below:

Event Detection/Trigger:

- Explicit manual trigger by a PS user (for a small geographic scope) or centrally by a PS entity interfacing to the network through a secure portal or device app.
- Automatically triggered based on monitoring and analytics-driven analysis of PS usage, network conditions, and environmental sensor data.
- Combination of above.

PS Prioritization Actions (within affected area):

- Dynamically raise the QoS priority level of PS users.
- Dynamically change the call admission control (CAC) policies to only admit PS users onto the wireless network (PS shared only, or a broader set of bands).
- Dynamically reserve a significant slice of RAN resources (by spectrum band, or by resource blocks within bands) for PS use.
- Increase capacity of specialized PS vEPC (virtual packet core).
- Combination of above.

6.4.3 Input Data

The types of inputs that could be collected and analyzed to determine when and where PS priority must be invoked falls into the following categories:

Explicit requests – An individual PS user with an immediate tactical emergency need, many PS users simultaneously requesting (indicating a broader emergency need), or a PS entity centrally requesting PS priority to deal with a large scale emergency situation.

PS user experience performance – Monitoring of KPIs for quality of PS connectivity relative to critical PS application(s) need for individual PS users, to identify situations where needed emergency access to the network (despite high priority access class) is blocked by congestion, network failure, or any other reason.

Network congestion – Monitoring of network congestion by location (per cell), in order to correlate with PS user experience KPIs to identify insufficient PS access.

Network failure – Widespread (multi-point) network failures or degradations which may be an indicator of a disaster situation, and may also be limiting PS access to the network.

Environmental indicators – Data from network hosted and other infrastructure sensors which collectively would be indicative of a disaster event. This could include fire/smoke detectors, vibration sensors, power loss sensors, infrastructure (i.e., bridge, building, and tower) stress sensors – as well as commercial and residential alarm sensors.

6.4.4 Output Action & Target

Once it is determined that priority access needs to be invoked for PS, actions can be directed toward the mobile core through PCC policy mechanisms, toward the RAN for direct control over air interface resources, or both. Notifications toward users and to network operations staff would also be important meta actions. Examples are outlined below:

Core/Policy Targeted Actions – Remapping PS users to a higher priority class, increasing PS maximum throughput per session, and/or limited non-PS throughput.

RAN Targeted Actions – Change CAC policies to enhance PS users access, or limit non-PS user access to one or more spectrum bands; adjust RAN scheduling algorithm to reserve priority access to a portion of RAN air interface resource blocks for PS users.

Transport Infrastructure Targeted Actions – Prioritize PS traffic through the backhaul network by raising the DiffServ Code Point (DSCP) and/or Layer 2 priority marking; initiate reconfiguration/re-optimization of network (L3-L0) connectivity to critical PS wireless communications sites (vEPC DCs) - as well as other critical PS sites (police, fire, emergency management centers, etc.).

Notifications – Informing PS users of their priority access status, other users of the reason for potentially degraded service, and network operations staff of the extent of PS priority being invoked.

6.4.5 Data Correlation Needs

Data correlation needs would include:

- To detect tactical PS priority needs, the usage and attempted usage of individual PS users would need to be correlated with the PS-relevant applications they are using as well as the network congestion levels of the affected cell(s).
- Automatically detecting the need for disaster-level PS priority involves geographic correlation of a wide range of (1) PS users' performance and priority requests, (2) network failure conditions, (3) environmental sensor data, and (4) other alarm sources.

6.4.6 Gaps

A use case such as this, pulling data from many different types of network and non-network sources together for analytics-driven correlation, will need to leverage a building block approach that breaks the problem down into more manageable analytics tasks. This will likely also be a necessity in order to leverage existing systems which may already perform related functions. Building blocks might focus on user experience monitoring, network failure correlation, external sensor network correlation, etc.

In order to perform the higher layer uber-correlation function, standardized mechanisms for sharing the output of the lower level analytics will be needed. These will need to include geographic information, which may not be an explicit part of all of the sub-systems data models today.

Control over sliced RAN resources, either directly or via Network Management or SON systems, is another area likely requiring new standards.

6.5 Use Case 5: Enhanced Fault Resolution

6.5.1 Description & Business Value

IP Service delivery can involve an increasingly complex collection of network functions (e.g., content filters, caching, CDNs, and content optimization proxies). Services can take various paths across the network, use a variety of interconnection points between networks, and access a given service from a number of different web servers. This can lead to wide variations in service experience, with no apparent differences in configuration when viewed from the user's perspective. When there are network problems, either due to transient congestion or due to incompatibilities between specific equipment types, identifying and correcting the problem can be extremely difficult. From the user's perspective these appear as intermittent or random problems, making them harder to isolate and more frustrating to deal with.

Collecting and correlating data from devices, even within a single network, would allow the network provider to identify that there is a problem before the user even realizes that there is a consistent problem. As complex services combining components from multiple entities continue to grow, resolving this sort of problem will become more and more of an issue.

The business value from this capability can be viewed from several perspectives:

- The ability to provide better user QoE;
- More effective fault isolation, reducing maintenance costs and negative publicity; and
- A service that could be offered to application and content providers to help them.

6.5.2 Variations

The correlation of data for a given service could involve the following alternatives:

- Identify all the equipment involved throughout the network for a given service;
- Potentially extend this to identify equipment outside the network, such as interconnection points, CDNs or even specific content servers. (NOTE including functions outside of the service provider network may be out of scope for this analysis.);
- On request, collect version and configuration information from each function involved in the end-to-end provision of a given service;
- On request, collect performance statistics (e.g., buffer fill, latency, dropped packets, etc.) from each function involved in the end-to-end provision of a given service. (NOTE: this would require the

ability to determine the path across the network, along with an accurate timestamp to allow correlation of information.)

6.5.3 Input Data

The types of inputs that could be collected and analyzed include the following:

- Information about the data path across the network;
- Equipment configuration and version;
- Congestion status;
- Packet loss; and
- Interconnection points to other networks.

6.5.4 Output Action &Target

Once it is determined that information for a given service is to be collected, all information associated with that service across the network would be collected, and forwarded to one of the following:

- A fault location application within the network; or
- A fault location application associated with the external application or content provider.

The information collected would have to include sufficient detail (e.g., timestamps) to allow adequate correlation.

6.5.5 Data Correlation Needs

Data correlation would need to do the following:

- Correlate congestion and packet loss data end to end to identify where the source of the problem is; or
- Look across different instances of similar problems to identify the common elements associated with the user problems. This could be specific interconnection points, equipment, versions of equipment, or combinations of equipment/versions. This implies the ability to collect and retain detailed information each time a problem is reported (potentially some time after the problem occurred) and look at the detailed configuration information to identify patterns.
- The ability to identify specific NFV instantiations involved in a service path and the ability to instantiate monitoring functions.

6.5.6 Gaps

A use case such as this, pulling data from many different types of network and non-network sources over an extended period of time for analytics-driven correlation, will need the ability to store network information along with accurate timestamp data to allow correlation. It will also require detailed information on the equipment, such as equipment model, version, and software release. The techniques identified in previous use cases, including the ability to define and combine modular analysis, would also be applicable. NOTE: Include potential NFV gaps as well.

6.6 Use Case 6: Outage Alerting, Avoidance, & Reporting

6.6.1 Description & Business Value

As the PSTN and wireless networks transitions to all-IP communications, many aspects of the way networks are managed must also change, presenting new challenges. One such challenge is how regulatory-relevant outages are measured and reported in IP networks, and how faults that cause reportable outages can be identified for reporting and restoration activity.

On February 15th 2012, the FCC approved a report and order expanding the mandated outage reporting rules to include VoIP telephony services, effective December 17, 2012. The new VoIP requirements are similar to current voice regulations today for Circuit Switch telephony; specifically the following outage conditions require reporting:

- Telephony User Outage Condition: >=30 minutes duration and 900,000 user-minutes.
- Special offices: airports – direct services to airport operations: >=30 minutes duration.
- E911 rules redefined (>=30 minutes duration and >=900,000 user-minutes and (1) PSAP loss of communication or (2) E911 loss of call processing (tandem, router, switch, etc.) or (3) loss of ANI/ALI.
- Mass Blocked call events that result in >=90,000 real-time blocked calls and last >=30 minutes.
- Further complicating factors are that both network operators and OTT providers are required to report VoIP outages, and the FCC requires localization information for outages.

In the PSTN, discrete voice switches and voice lines and trunks of deterministic voice call capacity makes the determination of the number of lines impacted by any given fault (e.g., a switch port card failure, or TDM trunk failure) fairly straightforward. By contrast, IP networks are converged service networks where voice traffic typically represents a small proportion of the aggregate traffic through any given link or switch, and the throughput per voice connection varies significantly and continuously over time. In addition, IP congestion control mechanisms, depending on how the network is engineered, may not fully restore impacted voice connections. So the question arises: how to achieve the level of visibility and control needed to both accurately measure and to minimize network outages in IP networks?

The goal of this use case is to improve outage-causing fault visibility, control, mitigation, and even prevention capabilities via automating the collection and analysis of key information gathered from across the network, as well as alerting and automatic network/service restoration actions.

6.6.2 Variations

There are multiple levels of scope and sophistication of this use case that can be contemplated, for example:

- Automated outage threshold detection – tracking locations of faults, calculating the total number of users impacted, and tracking the duration of those faults in order to identify when thresholds have been crossed for an outage requiring reporting, as well as aggregating the total user voice minutes impacted by an outage.
- Automated early outage warning – an early warning version of the base variation, which would alert network operations staff of network faults that threaten an impending reportable outage, and the projected time available before the threshold is crossed.
- Automated network repair/mitigation – taking the previous variation to a significantly higher level, it may be possible under certain types of network faults to employ automated instantiation of new virtual network functions and/or rearrangement of network connectivity in order to eliminate or reduce the outage impact of a network fault. For example, if a large scale power or connectivity failure made local Media Gateway (MGW) or Session Border Controller (SBC) elements inaccessible, new vMGW or vSBC capacity could be brought online quickly at another data center location and VPNs could be reconfigured under automated SDN control to route VoIP traffic to that replacement virtual network function.

6.6.3 Input Data

The types of inputs that would need to be collected and analyzed to determine the magnitude of user-minutes outage impact due to network faults in the network falls into the following categories:

Application-specific traffic monitoring – Monitoring of key data plane interfaces/links for the volume of voice traffic being transported, which may be a subset of the traffic being carried.

User-specific traffic monitoring – Monitoring of the traffic of special classes of users (e.g., airport operations, TSP) and special types of calls (i.e., E911), require unique reporting requirements.

Fault alarms & metadata – Collection of primary (point of fault) NE and link alarm information (both general and E911 related), localizing the elements and interfaces impacted.

Session drop statistics – Monitoring of voice session call control and/or data plane (e.g., Session Border Controller) connection points for abnormal call termination and failed call attempt statistics.

Outage localization – If not available directly from alarm information correlation, localization could be approximated by comparing current vs. historical voice traffic patterns to estimate the geographic reach of an outage.

SIP registration statistics - Registration failures at various points in the network can be used to identify outage conditions based on users affected.

Network VoIP statistics – Tracking of the total numbers of voice lines on a network (assigned numbers, active, inactive,)at various points would aid in determining outage impact for events where access to such data is unavailable. The FCC's Disaster Information Reporting System (DIRS) requires total VoIP Subscribers Served vs Total Subscribers Down be reported by county served.

6.6.4 Output Action & Target

Once a network failure impacting voice traffic is detected, and the magnitude and extent of the voice traffic impact is assessed by the analytics systems, a number of alerting and reporting (and possibly even mitigation) actions need to take place. Examples are outlined below:

Voice-impacting fault alert – Operations alert highlighting network fault(s) that have the potential to cause a reportable voice outage, so that service restoration can be prioritized.

E911-impacting fault alert – Operations alert highlighting network fault(s) that have the potential to cause a reportable E911 outage, so that service restoration can be prioritized.

Special user-impacting fault alert – Operations alert highlighting network fault(s) that have the potential to cause a reportable Special User outage, so that service restoration can be prioritized.

Reportable outage threshold crossing alert – Operations alert indicating that network fault(s) have resulted in a reportable E911 outage, so that service restoration can be expedited and collection of outage reporting information can be triggered.

Outage summary report – An automated report generated to summarize the information needed for an outage report, either after services have been restored or ongoing "in-progress" reporting of cumulative outage statistics.

Mitigation actions – Beyond normal protection mechanisms, an SDN-controlled and/or NFV-implemented infrastructure may be triggered to re-optimize the network connectivity and placement and scale of VNFs in order to fully or partially mitigate the impact of an outage. The range of potential network actions is broad, and beyond the scope of this document to enumerate.

6.6.5 Data Correlation Needs

Assessing the extent and localization of outages will require extensive correlation of traffic data and network information culled from a variety of sources across the network, and may vary significantly depending on the level of information already correlated in OSS systems employed in any given network. The following are examples of the types of data correlation needed to drive this use case:

Timestamping – accurate timestamps for all network alarms (on and off), IP traffic monitoring samples, and session traffic statistics.

VoIP address localization – mapping of VoIP traffic (or call attempt) IP addresses to dataplane control point (SBCs) and/or user endpoint locations.

Alarm correlation – reduction of multiple related alarms to primary (causal) failures, so that each network failure is only counted once for outage reporting purposes. Also correlation of node and link alarms to before and after traffic monitoring data (i.e., correlation to the failure group size – card, node, link, etc.).

6.6.6 Gaps

- Hardware component alarms generate the need for repair activity. Duplicate or additional inquiries denoting FCC outage threshold have been met are not desirable for non-outage conditions, i.e., if rerouting for traffic convergence has occurred.
- Outage reporting rules require that outages be defined in geographical areas at the county or community level. In an all IP environment, at what geographic area can similar impact be identified?
- Session dropped call metrics need to be defined to include industry standard error logs that can be used across wireline, wireless, cable, and OTT providers that can be consistently and reliably monitored.
- Alarm points need to be monitored for outage conditions and need to be further defined.
- Identification of black hole conditions (i.e., outage conditions that do not alarm) that give the impression that traffic is routing normally, but voice packets are being dropped.

6.7 Use Case 7: Network Wide Intrusion Detection

6.7.1 Description & Business Value

Security provides the foundation of service assurance. Miscreants and the threats that they impose against the networks used to deliver critical services continue to get smarter, more agile, and more destructive. Networks used to deliver applications continue to converge, making it more important to properly segment threats and vulnerabilities by domain, but examine the aggregate threat landscape at the same time. Examples of this include the mobile network evolved packet core (EPC) where traditional and mobile services share an infrastructure leveraging the Carrier Data Center and Cloud for operational efficiency and service delivery.

A network that uses NFV and SDN to deliver elastic services is typically secured using traditional security Best Current Practices that have been deployed into Service Provider networks for many decades. There is, however, an increase in the depth and breadth of the threat surface introduced by one of the key advantages of NFV and SDN: agility and dynamic deployment of services. Architectural innovation introduced by SDN introduces a new set of threats and mitigation strategies and also introduces a new set of visibility and control elements to handle the evolved threats. The agility and efficiency that NFV brings to secure the infrastructure is delivered when the system can mitigate a threat as close to the source as possible with the most appropriate control. Delivery of the right security mitigation based on innovation in how network information is seen and used via analytics and analysis deploys the workflow (series of mitigation controls) required to fix the issue dynamically. This is the case whether the workflow is to be deployed in a fixed line environment (into the correct VRF for the target in question) or into a mobile environment (offering the security services in the GiLAN or protecting services in the EPC).

Because no two threats are the same, the ability to apply the security VNF's dynamically at the right place in the network (based on a clear knowledge of the threat and how it impacts the downstream topology) is a "must have" in network operations today and tomorrow. In order to properly secure the full service environment that delivers a connected application, two fundamental elements are applied: visibility and control.

Visibility and Control together is the foundation for secure NFV. The strengths of SDN from the security perspective are:

- Easier administration of security mechanisms for virtual environments;
- Controls move with VMs without requiring manual intervention or reconfiguration;

- Integration of the virtual network and the physical network; and
- Removal of the dependency of the virtual network on the physical infrastructure.

At the same time SDN brings new risks from security perspectives which include:

- Control is through the virtual environment not through physical security mechanisms leading to potentially lower assurance;
- Management systems for SDN environments are immature and unproven from a security standpoint;
- No single dominant approach to implementing an SDN exists leading to a requirement for multiple types of security and management controls;
- Role separation between network, security, and server management is combined into the role of virtual environment administrator; and
- The Service Provider IT organization may be impacted by combining roles.

6.7.2 Variations

When it is put together, there is a need for a network wide policy context that is enforced at control points with VNF's (virtual functions providing a workflow) in a scalable and elastic manner. This is a critical operational behavior as now the security VNF's like the next generation IPS, Firewall, and DDoS service nodes (just a few examples) see threats in real time (behavioral or signature based) and programmatically redirect flows as appropriate. This closed loop feedback capability is only possible with an "end to end" perspective but specific implementations can vary based on network architecture and topology. Input data can be gathered from a variety of sources and the resulting mitigation can be provided using a variety of security devices and mechanisms as discussed above.

6.7.3 Input Data

Visibility refers to the ability to see and correlate information from the network as a whole to baseline proper behavior and then to measure deviation from that norm. Sources of visibility come from traditional network measurements (DPI knowledge, netflow, open flow, flow records …), or device measurements and triggers. Regardless, there is advantage in measuring all aspects of a flow, from all elements of the network to the application/device used by the end customer. An example of applicable input data include:

- Use of application level probes that are synthetically generated and travel through the network to get a clear picture of how an application is behaving.
- Path Computation Elements, which typically have a near real time database representing the network topology, and can be queried programmatically to determine the impact of a potential mitigation action on critical service classes and be used to specifically identify where additional security appliances might be instantiated as part of the mitigation action.
- Network analytics can applied to deliver a baseline of the network in good working condition and then to highlight when an anomalous event occurs. Thresholds on specific metrics or combinations of metrics can be used, based on history, as a signature to trigger mitigation actions. For example, exceeding a threshold may cause a security monitoring function or probe to be inserted at the appropriate place in the network to get more information.
- Device based security probes can notify the network of anomalous traffic patterns on the device.

6.7.4 Output Action & Target

Output action or "control" refers to the actions taken to mitigate an attack. Broadly speaking there are two major types of attacks on networks:

- Day zero attacks are threats where there is no available fingerprint (behavioral detection).
- Day one attacks are threats where there is a signature or fingerprint for and, quite often, a mitigation strategy exists in advance to handle the attack.

Once all of the telemetry/analytics is gathered, a network security controller can analyze the data and determine, based on policy, suggested mitigation and controls to be applied. Examples of mitigation actions would include:

- Use of NFV Orchestration and SDN to instantiate appropriate firewall, IPS, and DDoS service nodes into the flow path as necessary.
- Provisioning of Firewalls or ACLs to address specific attacks.
- Redirection of traffic to portals or delivery of alerts for end customer notification.

6.7.5 Data Correlation Needs

Data correlation needs include:

- Correlation of any device triggers with the location of the network security services that could be dynamically inserted. This may be complicated by the use of overlapping private IP address spaces where the IP address of the device is no longer unique.
- Correlation of network flow anomalies with other device/subscription information.

6.7.6 Gaps

Gaps associated with this use case are generally related to the data correlation needs of the previous section.

6.8 Use Case 8: NFV Automated Network Growth/Degrowth

6.8.1 Description & Business Value

Capacity planning is a critical function within the Service Provider network. Traditionally, dedicated hardware based deployments require long lead times to engineer, order, install, and provision new upgrades as the network grows and changes with dynamic user traffic patterns. Given the long interval required, capacity planning activities must estimate future needs based on historical usage as well as anticipated and known trends in traffic patterns and application usage. This inevitably results either in:

- Significant over provisioning with an associated capital and operations costs.
- Under provisioning that may result in impaired user experience affecting churn rates and jeopardizing revenue growth.

The elasticity associated with NVF and SDN enabled networks goes a long way in mitigating these negative consequences by both automating the grow/degrowth process and enabling much shorter time-to-turn-up to better utilize network assets.

In all cases, traditional network performance tools that monitor network transport and nodal capacity / utilization are used. These tools can provide a historical view of the time varying capacity and performance attributes of the network. By analyzing these data feeds and driving the results into and NFV orchestration system, automated changes in network capacity can be achieved.

6.8.2 Variations

In addition to traditional performance management metrics, NFV specific systems within the NVF infrastructure can provide additional information regarding Virtual Network Function CPU, memory and network utilization metrics that can used in the automated growth/degrowth process. Additionally, SDN based systems can resize transport links to manage overall transport network efficiency.

6.8.3 Input Data

Input data would include traditional performance management metrics including:

- Per transport link (e.g., MPLS path and/or physical link) packet loss and link utilization.
- Per network element (VNF) CPU, memory and storage utilization as a percentage of maximum along with application specific metrics with appropriate thresholds.
- Per network element bandwidth utilization on interfaces as a function of the maximum bandwidth
- Network link and node failure status and redundancy conditions.

6.8.4 Output Action & Target

Specific recommendations to increase/decrease capacity would be provided to an NFV/SDN orchestration element. This function would orchestrate the required actions to spinup/down Virtual Machines to increase/decrease VNF capacity as necessary. Infrastructure SDN capabilities would be used interconnect the new instantiated VMs into the system and to resize any WAN links to accommodate traffic demands as needed.

6.8.5 Data Correlation Needs

Network topology data is required in order to properly correlate node capacity needs with transport link capacity needs to properly size the system.

6.8.6 Gaps

This use case is being addressed by ETSI NFV and other standards fora and as such, no standards gaps are anticipated at this time.

6.9 Use Case 9: Inter-DC Congestion Migration

6.9.1 Description & Business Value

Web applications are largely implemented in and across cloud data centers, enterprise WAN networks increasing interconnect enterprise Data Centers (DCs), and networks will become increasingly implemented as virtualized network functions distributed in many parts of the network. As a result, traffic between data centers represents a rapidly rising and mission critical component of web services, operator services, and enterprise operations.

Inter-DC traffic patterns are highly dynamic by nature, as very large flows kick in sporadically as virtualized workloads are scaled or shifted/rebalanced between sites, applications perform geo-redundancy synchronizations, and restoration events redirect around failures. Many of these large flows are sensitive to transfer duration and latency. Overbuilding intra-DC links, on a conventional static basis, to ensure high performance for such dynamic peak demands is economically impractical. However, with SDN flexible programmability extended beyond the DC to the broader WAN, inter-DC connectivity can be dynamically and elastically rearranged in response to these rapidly varying demands.

Network intelligence can be leveraged to detect when inter-DC demand patterns have shifted, causing congestion, or even to proactively predict when such shifts are likely to occur. By leveraging a network-wide view of the network resources and utilization levels, SDN control can be used to modify the capacity matrix of the inter-DC network to better fit the network to immediate demands and minimize congestion impairments due to demand spikes. This also opens the opportunity for new elastic connectivity service models analogous to (and complementary to) popular cloud elastic compute and storage models.

6.9.2 Variations

While there can be many forms of this use case, a couple of significant dimensions of variation are the following:

Application coordination – whether the network (a) makes use of direct interactions with applications to request and/or schedule express capacity for large flows, or (b) relies entirely on network intelligence available within the network.

Whole or partial elastic inter-DC network – whether (a) the entire inter-DC network is reconfigured/re-optimized as demands vary, or (b) a specialized overlay portion of network capacity (perhaps engineered for very low latency) is created for the purpose of expressing large flows.

6.9.3 Input Data

The types of input data needed for this use case fall into the following categories:

Current network state – the topology and current link utilization of the inter-DC network (networking layers 3 to 0). This type of information would be available from centralized SDN multi-layer resource management functions, necessary for general WAN network programmability.

Explicit demands – Some new demands coming on-line (from new VNFs being instantiated, or a VM migration) may indicate their bandwidth and QoS needs explicitly, and this information can be used directly to anticipate congestion or the inability to meet an explicit service level agreement (SLA) if it is served by a managed connection service.

Large flow detection – If extra/express capacity is not explicitly requested by an application, such large flows would need to be detected by network monitoring of traffic between data centers. This might be accomplished by 5-tuple inspection, deep packet inspection (DPI), or selective DPI monitoring of a pre-classified subset of traffic.

Congestion detection – Utilization thresholds on inter-DC links need to be continuously monitored, so that congestion condition alerts can be triggered, and also so that future congestion conditions can be projected.

Historic demand trends – Regular sampling and storage of inter-DC traffic by time of day and application type information. This historical information would likely be aggregated and correlated by a network traffic analytics subsystem, which would in turn make it available for the analytics application implementing the inter-DC congestion mitigation algorithms.

6.9.4 Output Action & Target

For requested or scheduled express flows onto special express paths, the following are examples of the types of outputs are needed:

Express capacity advertisements – “green light” indications of the availability of express capacity to other data centers, signaled from the network to applications.

Express flow grants – the response to application requests for express capacity, either confirming (granting), modifying (countering), or denying the request based on available express capacity.

For both requested/scheduled and purely network intelligence-driven optimizations, the following types of outputs are needed:

Flow redirect to express path – rerouting of an existing flow or pre-selection of a future flow to a special express path. This may be accomplished via classic OpenFlow control, setting flow routing within a flow fabric under direction of an SDN controller, based on inspection of the first packets in the flow.

Network re-optimization trigger – When congestion in the inter-DC network is detected or is deemed to be imminent, a trigger is required to initiate a re-optimization of the L3-L0 layers (or a subset) of the network. This would, for example, cause a centralized SDN resource management, or an optimization application acting upon it, to change the multi-layer capacity matrix (e.g., adjust

the link capacities on the router links by changing optical transport connectivity) in order to better match the current traffic patterns and hopefully eliminate the congestion.

6.9.5 Data Correlation Needs

New demands – requested, detected, and/or projected - need to be correlated with the existing inter-DC network capacity matrix and current utilization in order to determine the appropriate flow steering and if further network (re)optimization is required. For flow steering, identifiers (classification information) for new flows need to be bound to express path identifiers.

6.9.6 Gaps

While standards and APIs for general multi-layer SDN control and resource management will continue to evolve and mature, one area for development specific to this use case include API interactions between applications and the network for express capacity advertisements, requests, and grants. These may vary from standard APIs requesting connectivity, in that they involve special QoS requests on top of existing connectivity. Another somewhat more general area is common APIs to expose historical usage pattern information from network analytics.

6.10 Use Case 10: RAN-Aware Time Shifted Content Delivery

6.10.1 Description & Business Value

Under RAN user-plane congestion service providers may elect to defer delivery of certain services, e.g., less time-sensitive traffic, to a later time when the network load is lower.

The goal of the RAN-Aware Time Shifted Content Delivery is to enable Service Provider to deliver managed content to a user at times and locations where the network is uncongested leveraging RAN congestion awareness in order to:

- Enable service providers a means of better monetizing their excess capacity.
- Encourage users to accept delayed delivery via charging incentives.

Consider for example a Service Provider offering 3rd party/OTT services to their users. The SP and the OTT operator have SLAs with regards to the sponsors, QoS, charging, revenue sharing, maximum number of simultaneous connections etc. The SP may provide incentives to 3rd party/OTT to accept differed delivery of content when the network is under congestion.

Another example would be a SP offering incentives to individual subscribers to defer delivery of content for certain applications when the network is under congestion.

6.10.2 Variations

- Extend delivery and RAN awareness across multiple access layers and/or technologies.
- Extend to non-managed content via a Network API.
- User QoS profile extensions.
- OTT Provider QoS profile.
- Enriched analytics.

6.10.3 Input Data

- OTT/3rd Part Application domain:
    - OTT priority.
    - Application-Id.
    - DeferredDelivery-Indication.

    - Mobility data.
- SPR/UDR:
    - OTT provider profile.
    - Extended Subscriber QoS profile.
- Cell Congestion Level.
- Mobility status (i.e., whether the subscriber is moving and is subject to frequent handovers).

6.10.4 Output Action & Target

RAN-Aware Time Shifted Content Delivery could be facilitated if the information elements (IEs) listed in the previous section are available at the PCRF.

Once the PCRF decides that delivery of content will be deferred then it performs the following function:

1. Determines the re-try interval based on the RAN congestion level.
2. Derives and stores in its own database or in SPR “DeferredContent-ID”.
3. Sends a reply to the application function that includes the retry-interval, “DeferredContent-ID” DeferredDelivery-Indication.

When the re-try interval expires the AF initiates a connection to the PCRF it includes the DeferredContent-ID and DeferredDelivery-Indication IEs. The PCRF determines the PCC rules that include the ChargingKey and the DeferredDelivery-Indication and provisions the policies at the P-GW/PCEF. The PCEF includes DeferredDelivery-Indication IE in the Gy message it sends to the OCS. The OCS provisions credit-rules at the PCEF taking into account the DeferredDelivery-Indication. The DeferredDelivery-Indication IE is also sent to the OFCS over the Gz interface. The PCEF includes DeferredDelivery-Indication IE in the CDR.

6.10.5 Data Correlation Needs

To properly implement this use case, certain aspects of the input data need to be correlated. For example:

The PCRF needs to correlate:

1. DeferredDelivery-Indication in the OTT’s and UE’s QoS profile in the SPR must be correlated with the Rx IEs.
2. The OCS must correlate the subscriber-ID and the DeferredDelivery-Indication IE.

6.10.6 Gaps

PCC Rx and Gx support for DeferredDelivery-Indication and DeferredContent-ID.

- P-GW and Gy/Gz (OCS/OFCS) support DeferredDelivery-Indication.
- GSMA OneAPI support for DeferredDelivery-Indication.

6.11 Use Case 11: Dynamically Inspect Traffic & Predict Performance of the Network

6.11.1 Description & Business Value

The objective of the use case *Dynamically Inspect traffic and predict performance of the Network* is motivated by two factors. One is the proliferation of bandwidth intensive applications such as increasing use of streaming multi-media per household and further, simultaneous use of multiple user devices (e.g., Internet enabled TV, use of PC for OTT multi-media services and Wi-Fi use of smart phones for viewing multi-media). Two, broadcast TV offering now increasingly includes IP based delivery of video services, interactive games, and IPTV. Both of above factors in combination are continuingly placing an ever increasing demand on the service delivery network. Communication Service Providers’ networks support delivery of these bandwidth intensive multi-media applications over its evolving infrastructure; however,

besides the obvious self-expectations of delivering content without delay, jitter, and packet loss CSP's are increasingly putting a heavy emphasis on being able to proactively monitor the network for quality of service delivery including the ability for predictive monitoring of the network to help improve quality of customer experience. The ability to learn traffic patterns based on storing customer centric network performance (KPI's and health of network), its use by application type can be used to detect and predict network performance degradations, network failures, and future performance of the network based on changes in conditions and traffic.

6.11.2 Variations

There are additionally possible uses for this type of application intelligence. For example:

- Generate reports on multiple level (traffic type, destination, and application).
- Ability to know how traffic is routing at the peering points and within the core network at any given point in time.
- Ability to monitor and trend impacts from changes in the network.

6.11.3 Input Data

The following collection of subscriber based usage (individual and aggregated); network usage per application, and aggregated network utilization would be useful for this use cases. This includes but not limited to:

- Broadband data usage per subscriber (e.g., Bandwidth usage per hour; per day; per month, yearly, etc.).
- VOD traffic per subscriber (per hour, per day, weekly, per month, annually).
- VOD user interaction with remote control and response time for actions such as pause, resume, play, browsing programming guide, and user transactions such as buy, rent or retrieve prior rental or purchased items.
- VOD traffic at Central Office (CO) and upstream to VOD servers and infrastructure in the middle including Policy Managers.
- IP TV usage per subscriber, aggregated subscriber per CO.
- CDN volume to and from peering partners (increments of some duration).
- Linear video traffic per CO; jitter, latency, and packet loss.
- VoIP usage per subscriber; delay, loss of voice, quality.

Network data for this example could include:

- Subscriber IP address.
- Upload speed.
- Download speed.
- Packet Frame Loss Rate.
- Latency.
- Network Jitter.
- Aggregated network link utilization per CO and toward the core network (end user device, CO based devices, and upstream NE for applications such as VOD, IPTV, Broadband Service).

6.11.4 Output Action & Target

Application based actions for example is to collect customer centric network performance statistics as listed above in section 6.11.3. For example, above KPI's can easily identify local or broader network congestion or outages, and if so, have the ability to correlate outages to impacted customers and set up automated voice response with restoral time for customer notification until problem resolution or a temporary spike/congestion may not need a truck roll but data can be used for future capacity planning. Additionally, the metrics collected can be fed into the analytics engine for broader correlation such as proactively spotting trends and areas of possible congestions within the network or at peering points and

specifically learning the network usage pattern, degradation to help predict the future network performance.

To properly implement this use case, certain aspects of the input data needs to be correlated. For example:

- Correlate User IP address with WI-FI router and Set-top-boxes (STB) ID's (including all IP address inside the premise).
- Correlate user's IP and in-home devices to the network device outside premise.
- Correlate usage by application type. For e.g., correlate IP address and WI-FI router ID to broadband usage and likewise, WI-FI Router ID and STB ID's with VOD stream and IPTV etc.
- The per subscriber utilization can be aggregated to higher level such as by CO and also further aggregation up stream in the network and (aggregated subscriber per CO and aggregated link utilization) to help determine aggregated user subscription and network utilization by CO and upstream.

6.11.5 Gaps

The key gap associated with this use case is indicated by the data correlation needs of the previous section. Today, it is often difficult to correlate diverse data items that are often collected by systems that do not have the ability to associate the data with other critical information in the network.

7 SDO/Gap Analysis

7.1 Aggregate View of Use Case Inputs &Outputs

Many of the use cases described in Section 6 have information elements in common between them. In order to efficiently organize and focus the analysis of gaps, it is important to develop a summary view illustrating the common and unique data elements across the use cases. Table 7.1 summarizes the primary input data types involved in the use cases, organized by broad source category. Entries selected as data elements for further gap analysis are highlighted in **bold** with a reference to the specific gaps analysis table number **[7.2.n.m]**.

Table 7. 1 - Use Case Input Data Summary

Inputs			
Use Case	**Subscriber Data**	**Application-specific Data**	**Network Data**
1. Network-wide Application Detection and Usage Support	• Subscriber profile and policies.	• **Application Type [7.2.2.2].** • Per subscriber, per application BW usage and latency. • **Application Metrics [7.2.2.1]** - Per subscriber such as Video rates, video stalls, …	• Location. • **Congestion state along forwarding path [7.2.3.1].**
2. Prioritization of Traffic	• Subscriber profile and policies.	• **Application Type [7.2.2.2].** • OTT/3rd Application and connection priority.	• Location. • **Congestion state along forwarding path [7.2.3.1].** • Base station performance KPIs.
3. Personalized Broadband	• Subscriber profile and policies.	• **Application Type [7.2.2.2].** • **Application Metrics [7.2.2.1].** - Per subscriber, per application BW usage and latency. • By Time-of-Day.	

4. Public Safety Spectrum Sharing	• Explicit Request.	• **Application Type [7.2.2.2].** • **Application Metrics [7.2.2.1]**. - Per subscriber, per application BW usage and latency.	• **Congestion state along forwarding path [7.2.3.1].** • **Network Failure information [7.2.3.2].** • Environmental indicators.
5. Enhanced Fault Resolution	Device: • Type. • Address. • Configuration. • Version .	• **Application Type [7.2.2.2].** • **Application Metrics [7.2.2.1].** - Per subscriber, per application BW usage and latency. • OTT/3rd Application and connection priority.	• Equipment Location. • Network Interconnection KPIs. • **Congestion state along forwarding path [7.2.3.1].** • Aggregate packet performance metrics .
6. Outage Alerting, Avoidance & Reporting	• Identification of impacted special users. • Subscriber Location.	• **Application Type [7.2.2.2].** • App-specific (VoIP) traffic volume. • NG911 VoIP traffic.	• Network faults: **Utilizing SIP Error Codes to Identify Blocked Calls [7.2.3.3]**, and **Auto Detecting Out-Of-Service Conditions [7.2.3.4]**.
7. Network Wide Intrusion Detection		• Per subscriber malware detection triggers.	• Forwarding path information. • **Network performance metrics (per link, per node and interconnect KPIs).**
8. NFV Automated Network Growth/ Degrowth			• **Network performance metrics (per link, per node).** • **Network Failure information.**
9. Inter-DC Congestion Mitigation		• Express connectivity requests. • Large flow detection.	• Utilization/Congestion state. • Historical usage patterns. • **Network Failure Information.**
10. RAN-Aware Time-Shifted Content Delivery	• Subscriber profile and policies.	• Policies for OTT/3rd party application.	• **Congestion state along forwarding path [7.2.3.1]** (particularly RAN).
11. Dynamically inspect traffic and predict performance of the network	• Subscriber ID and performance metrics (IP address, upload/download rates, Packet Loss, Latency/Jitter).	• **Application Type [7.2.2.2].** • **Application Metrics [7.2.2.1].** - Per subscriber, per application BW usage and latency. • By Time-of-Day. • **Application Metrics [7.2.2.1]** - Per subscriber such as VOD user interaction, VOD guide usage, VoIP quality metrics …	• Aggregate Application (e.g., VOD) traffic per link including BW, packet loss and latency/jitter.

Table 7.2, below, summarizes the primary output data or action types involved in the use cases described in Section 6, organized by broad destination category. Entries selected as output/action data elements for further gap analysis are highlighted in **bold** with a reference to the specific gaps analysis table number **[7.3.n.m]**.

Table 7. 2 - Use Case Output/Action Summary

Outputs/Actions			
Use Case	**Output to User/App**	**Output to Operations**	**Network Action**
1. Network-wide Application Detection and Usage Support	• Confirm or delay QoS requests.	• Usage reports.	• **Application-specific Optimization Policy Decision/Enforcement [7.3.3.2].**
2. Prioritization of Traffic	• Confirm or delay QoS requests.	• Usage & network utilization reports.	• Traffic Prioritization (PEPs). • Optimization function parameters.
3. Personalized Broadband	• Personalized offers (SMS, IM). • Links to portal.	• Usage reports.	• Sub. DB update. • Update sub. charging & PEP policy.
4. Public Safety Spectrum Sharing	• Notifications of status to PS users.	• Ops alert of priority invocation.	• Traffic prioritization (PEP). • **Adjust RAN resource allocations [7.3.3.6].**
5. Enhanced Fault Resolution	• Notifications of service issue.	• Anomaly alerts. • Diagnostic data.	• Network and traffic stat correlation for anomaly detection. • Re-routing based on network condition.
6. Outage Alerting, Avoidance & Reporting	• Notifications of service issue.	• **Potential Outage Alarm Triggered by Blocked Called Events [7.3.2.1].** • **Alarm Indicating Auto Detected Out-Of-Service Conditions [7.3.2.2].**	• **Scale VNF capacity (auto mitigation) [7.3.3.3].** • Elastically scale network. • **Re-optimize L3-L0 network [7.3.3.4].**
7. Network Wide Intrusion Detection	• Quarantine app, device.	• Ops alerts. • Security event statistics report.	• **Quarantine device [7.3.3.1].**
8. NFV Automated Network Growth/ Degrowth		• Usage/utilization reports.	• **Scale VNF capacity (auto mitigation) [7.3.3.3].** • Elastically scale network.
9. Intra-DC Congestion Mitigation	• Capacity availability advertisements. • Express flow grants.	• Usage/utilization reports.	• **Flow redirection to express paths [7.3.3.7].** • **Re-optimize L3-L0 network [7.3.3.4].**
10. RAN-Aware Time-Shifted Content Delivery	• Delivered content.	• Usage/utilization reports.	• **Policy and Charging Rules for Deferred Content Delivery [7.3.3.5].**
11. Dynamically inspect traffic and predict performance of the network		• Detailed sub. KPI data analysis reports. • Detailed traffic/perf. trending reports.	

7.2 Gaps Analysis of Inputs (Information Elements, Triggers)

Gaps associated with selected entries from Table 7.1 are further analyzed in this section, categorized by column.

7.2.1 Subscriber Data

This category includes information about the subscription identity, policies, subscribed services/options, devices, preferences, etc. Subscriber data inputs identified in the Section 6 use cases where found to be generally addressed with existing standards associated with AAA, HSS, HLR, and DHCP network elements.

7.2.2 Application-specific Data

This category includes information about or from an application or application-specific information such as identity, traffic metrics, policies, etc. The data inputs which are both application- *and* subscriber-specific (which is typical in these use cases) are also covered in this section. Selected application-specific data inputs are analyzed in the tables below, referenced from Table 7.1.

7.2.2.1	**Application Metrics**

Key applications metrics, based on the use cases analyzed, include:

- Per application usage metrics such as bandwidth utilization, "tonnage" (number of bytes consumed), codec rates used (for media apps), time of usage (timestamp), time duration of usage.
- Application specific quality metrics such as packet loss, latency, jitter, video stalls, buffer fill, VoIP/media specific related quality metrics.

These metrics may be needed both in aggregate form (per link or node) as well as per subscriber.

Correlation Needs

1) For aggregate needs, the application metrics are needed typically on a per link and/or per node basis.
2) For per subscriber needs, the application metrics should be clearly associated with subscriber identity.

Importance

Any application-specific network optimization, diagnostics, charging, or other function requires the ability to detect and differentiate between applications, and to report to analytics engines application identifiers associated with flows in the network.

Source System/Interface	Current Standards Activity	Gap
DPI functions	3GPP has standardized a framework for DPI on the Gi and SGi reference points within the mobile architecture.	3GPP addressing the issues.
Network Probes	No known standards in place other than OAM related protocols and report formats typically used for monitoring/logging purposes.	None identified.
Device	Device can avail application specific metrics in a number of ways: • The device can provide such information as part of application specific protocol capabilities inband as part of the application operation. An example of this is RTCP used for RTP/VoIP communication metrics. • One can deploy a client within the device to collect application information and post this information to the network (or the application provider) out-of-band. This method typically would require an authenticated out-of-band connection.	May be a need for a standard out-of-band mechanism for reporting application specific metrics from a device to either the network or application provider.
Application Server	No known general standards to allow an application server to communicate application specific performance to the network.	None identified.

7.2.2.2	Application Type
Identification of the specific application associated with a given flow, detected at a designated ingress monitoring point in the network.	
Correlation Needs	
Needed for all of the applications metrics noted in the previous sections to correlate application information that may come from alternative sources (e.g., Network DPI, Network Probes, Device, and the Application Server).	
Importance	
Since application information can come from a variety of sources, a common standardized understanding of application type is essential when combining data from multiple sources. In cases where combining/correlation is done within the same service provider network, the service provider can enforce a common definition of application type. But in cases where application type is used across service providers or between an OTT and the service provider, a uniform definition of application type is needed. Commonly, a URI is often used as an application type identifier. However, the URI construct does not always lend itself to broader application type descriptions (e.g., ABR Video…).	

Source System/Interface	Current Standards Activity	Gap
All sources as noted in the Application Metrics table.	3GPP has done work within the PCC framework to standardize application IDs / types (see appendix).	Additional standards work is required to create a robust framework for common application type definitions.

7.2.3 Network Data

This category includes information about the network configuration, state, fault conditions, etc. which are not of a subscriber and/or application-specific nature. Selected subscriber data inputs are analyzed in the tables below, referenced from Table 7.1.

7.2.3.1	Congestion State Along Forwarding Path

There are two key aspects related to this information element, congestion status and knowledge of the forwarding path. Congestion status for most wireline elements (e.g,. links) is well defined today. Typically, metrics such as link occupancy (%), packet loss, and latency capture the essence of congestion in wireline. In wireless, congestion on the air interface is complicated by air interface impairments. For example, a specific subscriber may suffer from packet loss and latency due to poor signal conditions. The actual cell server the user may not be congested at all. Currently, the 3GPP RAN groups have not formalized a definition of congestion in the RAN. Such a definition would be useful in ensuring uniform interpretation of congestion indicators. Knowledge of the forwarding path is complicated by many factors. Within a specific routing domain, IETF standards have provided many mechanisms to aid in determining the forwarding path of a flow These include:

- ALTO – Application-Layer Traffic Optimization (RFC 5693, RFC 7285, RFC 6708).
- PCE – Path Computation Element (RFC 4655, RFC 6805).
- Segment Routing (particularly when coupled with a central SDN control to get metrics …).
- I2RS - Interface to the Routing System.

However, traffic may be tunneled via a VPN or for mobility reasons and may pass through different administrative domains making forwarding path collection complex if not impossible.

Correlation Needs

Congestion along the forwarding path must be relative to a specific flow or set of flows associated with a subscriber. Quite often, the user may know a link is congested based on occupancy and packet loss data but which subscribers are contributing to the congestion and how much may not be known.

Importance

This specific information element is common to at least 4 use cases as shown in table 7.1 above.

Source System/Interface	Current Standards Activity	Gap
Congestion information is typically available through the management systems managing the specific link/element. Forwarding path information may be available through a variety of IETF mechanisms as noted above.	3GPP RAN – UPCON work on user plane congestion. IETF activities include, ALTO, PCE, Segment Routing and I2RS.	No standard definition of RAN congestion in 3GPP. Need path information to be shared across different routing domains. Need capability to correlate congestion with subscriber identity.
Congestion can also be collected from network probes.	None known.	Potential need for standardized collection and correlation of congestion data.

7.2.3.2	Network Performance & Failure Metrics

Network performance metrics can be collected per link and per node. Network interconnect KPIs can be used to assess the performance of other interconnecting networks. Similar to congestion related information elements, network link metrics can include maximum link capacity and occupancy (%), packet loss and latency. Additionally, the network will know if a link or node has completely failed or not. In some cases, (for example on microwave links), the capacity of a link will vary based on environmental conditions. Network nodes can fail completely or partially creating lost packets and/or latency. Complicating network failure resolution is the fact that networks are layered with service specific nodes in a service layer, routing/switching nodes in a routing layer on top of a transport layer. The service layer can be comprised of elements such as VPN routers, CMTS/BNG nodes, mobility gateways, and other network appliances including firewalls and NAT devices.

NVF (virtualization) and SDN present both opportunities and challenges associated with collecting network performance and failure metrics. These new technologies may add new sources of failure that need to be detected, reported and fully integrated into the fault and performance correlation logic. However, these technologies can also make use of performance/failure metrics to dynamically improve performance and recover from faults through the network elasticity enabled by NFV/SDN.

Correlation Needs

All of network layers noted above have failure modes as well as recovery mechanisms that can create network performance anomalies. These anomalies must be correlated across layers to properly resolve a network failure or performance issue.

Importance

This specific information element is common to at least 3 use cases as shown in table 7.1 above.

Source System/Interface	Current Standards Activity	Gap
Network performance and failure metrics are typically available through performance and fault management systems. These systems include fault correlation features to best resolve network failures.	Basic standards in place for fault and performance data.	No gaps identified.

7.2.3.3	Utilizing SIP Error Codes to Identify Blocked Calls

In the legacy PSTN world, switches collected blocked call statistics that service providers used as one of many indicators to detect service interruption. Moving forward in the IMS world, some SIP Error Codes may be used as indicators for some types blocked call conditions. One function of Element Management Systems (EMS) or Service Assurance Monitoring tools is to collect and aggregate protocol error statistics occurring in the network. These systems can monitor SIP Error Codes between specific devices (i.e., between a Session Border Controller or IP interconnect device such as a MGW, used to interconnect to a carrier), within the local/market or throughout the entire network. Error count thresholds may be set to trigger alarms for critical and major blocked call conditions requiring incident investigation (refer to Gap Section below). The ATIS NRSC is investigating whether it is feasible to have additional thresholds set to identify potential FCC reportable events. This is contingent on a well-defined set of SIP Error Codes (new/modified) standards that may be used to identify blocked call conditions. A preliminary review by the ATIS NRSC revealed that SIP Error Codes do not always designate a blocked call condition; however, SIP Error Codes used in conjunction with a full signal trace may be used as an indicator of a blocked call condition. Some SIP Error Codes will require further refinement, additional filtering and correlation to indicate blocked call conditions.

Correlation Needs

The SIP Error Code data must provide a true indication of a blocked call. To avoid overinflating error counts, only the final response code needs to be taken into consideration. A baseline needs to be established for error codes expected during normal operations for the IMS network. Inbound error codes can be driven by several factors such as routing errors or cyber-attacks, neither of which are a true indication of a blocked call. The call blocking data must provide the ability to identify a geographic area of impact to aid in reporting and troubleshooting.

Importance

Identifying major blocked call conditions and providing alerting capabilities will minimize service interruption. This capability may also offer the industry a means to consistently and reliably report blocked call conditions for FCC outage reporting.

Source System/Interface	Current Standards Activity	Gap
Session Border Controllers / Element Management System / Service Assurance Monitoring tools.	ITU Standard. IETF RFC-3261 (SIP protocol aspects). 3GPP (IMS network level aspects).	Standard gaps: SIP Error Codes need to be further defined to accurately identify a blocked call condition. This assumes EMS/Service Assurance Monitoring tool can set thresholds and alarm. The reporting network elements have to report call completion rates or error code information regularly.

7.2.3.4	Auto Detecting Out-Of-Service Conditions

In an IMS environment there are many unique devices that make up the network. There are also multiple levels of redundancy or reroute capabilities engineered in the network. A failure of any one device or interface may not be indicative of an outage condition. So the question arises: how to achieve the level of visibility and control needed to both accurately detect and measure network outages in an IMS networks? Individual node alarms are required for break fix activity, but true failure conditions must also be identified so break fix activity can be appropriately prioritized. This will further reduce Time-to-Repair (TTR) intervals improving overall network availability. The ATIS NRSC is interested in using registration failures, Answer Seizure Ratio (ASR), and Network Efficiency Ratio (NER) statistics as a means of detecting true Out-of-Service (OOS) conditions. These statistics are collected by Session Border Controllers (SBC) or IP interconnect device such as a MGW, located throughout the IMS network. One function of Element Management Systems (EMS) or Service Assurance Monitoring tools is to collect and aggregate error messages/statistics across the network. These systems monitor between specific devices (i.e., between a Session Border Controller used to interconnect to a carrier), within the local/market or throughout the entire network. Thresholds may be set to trigger alarms for critical, major, and other OOS conditions requiring incident investigation (refer to Gap Section below). The ATIS NRSC is investigating whether it is feasible to have additional thresholds set to identify potential FCC reportable events. This is all contingent on a well-defined standard that can detect OOS conditions consistently. A preliminary review by the ATIS NRSC revealed that registration failures, ASR, and NER statistics do not always designate an OOS condition; however, these statistics used in conjunction with a full signal trace may be used as an indicator of an OOS condition. These standards would need to be further refined or would require additional filtering and correlation.

Correlation Needs

Registration failures, ASR, and NER data must provide a true indication of an OOS condition. ASR and NER statistics include Busy Signals, Ring No Answer, and Terminal Rejects which are not an indication of an OOS condition. To avoid overinflating the number of subscribers impacted additional filtering and correlation is needed. A baseline needs to be established to identify normal operating parameters of the IMS network. Statistical data must provide the ability to identify a geographic area of impact to aid in reporting and troubleshooting.

Importance

Identifying true OOS conditions and providing alerting capabilities will minimize service interruption and improve availability. This capability may also offer the industry a means to consistently and reliably report IMS voice OOS conditions for FCC outage reporting.

Source System/Interface	Current Standards Activity	Gap
Network Devices / Session Border Controllers / Element Management System / Service Assurance Monitoring tools.	Basic standards in place for fault and performance data. ASR - ITU SG2. NER – ITU E.411.	Registration failures, ASR, and NER statistical data need to be further defined, filtered, and correlated to accurately identify OOS conditions. This assumes EMS/Service Assurance Monitoring tool can set thresholds and alarm.

7.3 Gaps Analysis of Outputs (Control Points/Actions)

Gaps associated with selected entries from Table 7.2 are further analyzed in this section, categorized by column.

7.3.1 Outputs to Users & Applications

Output messages or signaling to users, their devices, or to applications as a result of the correlation, analysis, and decisions taken upon input data were analyzed for the listed use cases. Outputs/actions to users/applications were found to be generally addressed with existing mechanisms and standards. Specifically, the wide variety and availability of devices and user applications creates an environment where many application/device specific mechanisms have already been deployed using existing protocol standards. Given the diversity and implementation specific nature of these application specific

mechanisms, the focus group was not able to identify any specific standards changes necessary to implement the use cases analyzed.

7.3.2 Outputs to Operations Staff

Output messages or reports to network operations staff as a result of the correlation, analysis, and decisions taken upon input data were analyzed for the listed use cases.

7.3.2.1	Potential Outage Alarm Triggered by Blocked Called Events

If SIP Error Code standards are not created or modified to enable the Element Management Systems (EMS) or Service Assurance Monitoring tools to accurately detect and monitor blocked call conditions advance filtering may be required. The filtering, if needed, would enable EMS and Service Assurance Monitoring tools to analysis true blocked call conditions. The monitoring tools must have the ability to set thresholds and alarm on thresholding criteria (Critical, Major, Minor, and possible future threshold). A baseline needs to be established for error codes expected during normal operations for the IMS network. The output should be presented as an alarm with sufficient data to help isolate and troubleshooting the blocked call event and identify the geographic impact to customers.

Importance

Identifying major blocked call conditions and providing alerting capabilities will minimize service interruption. This capability may also offer the industry a means to consistently and reliably report blocked call conditions for FCC outage reporting.

Target System/Interface	Current Standards Activity	Gap
EMS or Service Assurance Monitoring tools.		No explicit output gaps, however; there is a need to align across the industry on best practices for analysis of network events to identify blocked call conditions.

7.3.2.2	Alarm Indicating Auto Detected Out-Of-Service Conditions

A preliminary review by the ATIS NRSC revealed that registration failures, ASR, and NER statistics do not always designate an Out Of Service (OOS) condition; however, these statistics used in conjunction with a full signal trace or a baseline may be used as an indicator of an OOS condition. These standards would need to be further refined or would require additional filtering and correlation. The Element Management Systems (EMS) or Service Assurance Monitoring tools must have the ability to set thresholds and alarm on thresholding criteria (Critical, Major, Minor, and other). A baseline needs to be established for normal operations for the IMS network. The output should be presented as an alarm with sufficient data to help isolate and troubleshoot the OOS condition, and identify the geographic impact to customers.

Importance

Identifying out-of-service conditions and providing alerting capabilities will minimize service interruption. This capability may also offer the industry a means to consistently and reliably report blocked call conditions for FCC outage reporting.

Target System/Interface	Current Standards Activity	Gap
Element Management Systems (EMS) or Service Assurance Monitoring tools.		No explicit output gaps, however; there is a need to align across the industry on best practices for analysis of network events to identify out-of-service conditions.

7.3.3 Network Actions

Automated network actions triggered as a result of the correlation, analysis, and decisions taken upon input data were analyzed for the listed use cases.

7.3.3.1	Quarantine Device

Quarantined devices could be subscriber devices or network servers being used to propagate malware. Generally speaking, a quarantine action would follow a sequence of events:

- Detection of traffic that is known to be indicative of the existence and propagation of malware from a source of traffic.
- Determination of actual end device identity (resolution of any network address translation or tunneling protocols).
- Redirection of device traffic to a quarantine server. This server may provide a web page providing instructions for the subscriber as to how to proceed in cleaning the device and resetting the network path. Additionally, the server could provide limited connectivity to the Internet (filtered for malware) to allow for some general Internet access until the device can be cleaned.

The redirection action can be facilitated through SDN control, through application specific mechanisms (HTTP Redirect) or via other mechanisms that enable source routing controls.

Importance

The ability to quarantine the offending device in Network Intrusion use cases enables effective action to be taken at the source of the intrusion.

Target System/Interface	Current Standards Activity	Gap
Network centric SDN Controller.	Industry wide SDN Controller use cases that allow for redirection of traffic to/from an IP address are already being considered in the industry.	No gap contingent on the outcome of SDN controller use case analysis.
Application specific mechanisms (such as HTTP Redirect) in combination with access control lists to block remaining traffic.	Typically well defined in IETF.	No gap identified.

7.3.3.2	Application-specific Optimization Policy Decision/Enforcement

A number of the use cases illustrated in the previous section leverage the ability to provide application specific enforcement actions. These actions can take the form of:

- QoS enablement where an application flow is given differentiated treatment.
- SDN based traffic management to route traffic through non-congested links and nodes.
- Adjustment of network parameters associated with application traffic to better optimize application behavior. For example, M2M devices (and associated applications) may be optimized by adjusting specific RAN parameters and timers to better suit application behavior.
- Routing flows to/through an application specific optimization function. For example, in mobile networks, TCP applications may benefit from TCP optimization capabilities.
- Providing application specific parameters to application optimization functions to enable more specific and granular optimization.
- Providing network performance and other information to application servers outside of the network operator domain.

Importance

Being able to enable application specific enforcement actions is a cornerstone capability that applies very broadly across user experience and network optimization use cases.

Target System/Interface	Current Standards Activity	Gap
Mobile Network Policy and Charging Control (PCC).	3GPP PCC allows for a variety of QoS and sponsored data features through the Rx interface. Additionally, the existing UPCON work item has identified a new interface into the PCRF for RAN related congestion use cases.	No gaps at present.
3GPP EPS / M2M applications.	3GPP continues to look at specific mechanism to optimize M2M traffic and applications.	No gaps at present.
Mobile (S)Gi-LAN.	3GPP has chartered an effort called Flexible Mobile Service Steering (FMSS) to study use cases and propose potential requirements for supporting traffic classification and service chain selection capabilities per operator's policy (e.g., based on user's profile, application type, RAN type, RAN status and flow direction) in order to realize efficient and flexible mobile service steering in the Gi-LAN network. This work could provide better application specific control mechanisms for mobile networks.	No gaps at present.
API Gateway.	External application servers may communicate with the network for application specific purposes.	No gaps identified.
Application functions within the network.	These application specific functions can pass information to external application servers by modifying or adding headers into the application header IP header fields. Typically, application standards are used to insure interoperability.	No gaps identified.

7.3.3.3	Scale vNF Capacity (for automated mitigation)

The ability to instantiate additional VMs for vNF capacity scaling, in some cases at an alternative site from the original serving capacity, in order to relieve functional congestion (overload), work around network congestion, or mitigate an outage.

Importance

The ability to leverage the dynamic scaling and placement flexibility of vNFs (instantiated through cloud management and orchestration functions) opens a whole new dimension for the types of adaptive network actions that can be taken in response to the analysis of network intelligence. The capability is a key enabler for several of the use cases outlined in Section 6.

Target System/Interface	Current Standards Activity	Gap
NFV MANO Functions.	MANO architecture and interface requirements are currently under development in the ETSI NFV ISG.	No gap contingent on the outcome of NFV ISG work, and assuming subsequent interface specification and information model standardization.

7.3.3.4	Re-optimize L3-L0 Network

The ability to dynamically reconfigure the “capacity matrix” of the IP/Ethernet/optical network, for example adjusting optical channel routing in order to shift capacity from less utilized router links to links needing to be scaled for purposes of relieving congestion, handling elephant flows between DCs, or rerouting traffic around outages.

Importance

The ability to leverage the dynamic programmatic network controls and network optimization algorithms so that the network can adapt to a variety of conditions is a major enabler for making the leveraging of network intelligence impactful. Most use cases dealing with (non-access) congestion or outage conditions benefit from this capability; indeed it is a key enabler for several of the use cases outlined in Section 6.

Target System/Interface	Current Standards Activity	Gap
Multilayer SDN Control system(s).	SDN Northbound APIs to “network controlling” applications are under consideration in various venues in the industry, including a new NBI-WG in ONF, but lag the development of Southbound (network-facing) interface specifications.	Expected to remain a near term gap, pending the successful completion of ONF NBI-WG work; not clear that additional parallel SDOs working NBIs would be productive at this point.

7.3.3.5	Policy & Charging Rules for Deferred Content Delivery

Many application specific network actions affect the policy and charging systems within the network.

Importance

Many networks, particularly in the mobility domain, have a rich set policy and charging rule capabilities to support application specific network actions. An example of these, relative to the deferred content use case from Section 6, include:

- Trigger content delivery of deferrable content.
- Set re-try timers and bandwidth limits in the application domain based on the congestion level of the network in response to the application function's request for QoS resources.
- Mark flows (Flow-Information) as "deferred" in a network database.
- Save application specific attributes such as a "deferred-indication" in the policy rule when deferred content is delivered.
- Discount charging when "deferred-indication" is sent to the charging systems from policy enforcement.

Target System/Interface	Current Standards Activity	Gap
Policy element (e.g., PCRF).	3GPP actively manages application specific policy and charging rule modifications.	Further investigation needed.

7.3.3.6	Adjust RAN Resource Allocations

The ability to explicitly control how Radio Access Network (RAN) air interface resources are made available to different groups of users, such as subscribers to different wireless service providers sharing the same RAN infrastructure, or users requiring preemptory access under certain conditions (e.g., public safety users during an emergency). This may involve one or more of several mechanisms for controlling access, including:

- Change call admission control (CAC) policies to shape the access of different user groups.
- Enable and disable access to different spectrum bands according to user group.
- Adjust RAN scheduling algorithm parameters to allocate or prioritize access – per user group - to RAN air interface resource blocks.

Ideally, these controls would make the wireless RAN programmable - analogous to the SDN-enabled programmability long envisioned and now emerging for IP and optical transport infrastructure.

Importance

RAN air interface resources are typically the gating resource for throughput capacity in any wireless network. Thus the ability to flexibly partition and allocate RAN air interface across different service providers and/or user groups will enable many new approaches and many new business models for efficient sharing of this fundamentally constrained resource.

Target System/Interface	Current Standards Activity	Gap
RAN Network Management and/or Self-Optimizing Network (SON) Control.	In the US, the Wireless Priority System (WPS) provides a standard mechanism for public safety personnel to receive prioritized CAC access via a special dialing code from a WPS-enabled device, mitigating the impact of congested mobile networks. The adjustment of RAN parameters, in areas such as CAC and air interface scheduling which directly impact base station operation, are typically controlled through proprietary interfaces from dedicated network management systems. While APIs to these management systems could be defined, the target mechanisms internal to the base stations are themselves not standardized, making standardization of external control interfaces of existing infrastructure very challenging. This may limit what is practically to achieve with current generation mobile networks to coarser level controls (e.g., binary access to different spectrum bands via basic on/off CAC). On the other hand, 5G standards are at a very early stage, and are expected to adopt more SDN-like approaches to programmability of wireless networks; this is an opportunity to define APIs upfront to enable standardized forms of programmable RAN control and optimization.	Open APIs to allow multi-vendor programmability by centralized network controlling applications to set policies for CAC and allocate RAN resources at Base stations between different service provider or user group “slices” (e.g., peak/committed rate or resource blocks, signaling channel capacity - per PLMN). Could be targeted at 5G standards, and adopted opportunistically by existing generation networks.

7.3.3.7	Flow Redirection to Express Paths

The rerouting of an existing packet flow, or pre-selection of a future flow, from the normal or default connectivity path to a special express path with the characteristics of higher throughput performance, better latency/jitter performance, and/or lower network cost. While potentially applicable to many networking venues, the most prominent use case involves inter-data center WAN connectivity for dynamic demands.
The flow redirection could be implemented in many ways, depending on the capabilities of the forwarding elements. One approach would be through the use of OpenFlow-controlled flow fabrics under SDN control, either establishing flow table settings a priori for anticipated or scheduled flows, or reacting dynamically based on inspection of the first packets in the flow and a policy pull via an SDN controller. Another approach would involve the use of Segment Routing by controlling (instantiating and updating) the path of SR-TE LSPs from a Path Computation Element (PCE), also in concert with policy-based overall programmatic SDN control. Service chaining-oriented traffic steering mechanisms could also be employed, although this specific WAN network oriented use case involves underlay as well as overlay networks, and does not require the same level of flexibility as service chaining.

Importance

The ability to flexibly steer individual flows in the network is a foundational component of building SDN-enabled programmable networks. The redirection of flows to an express path is just one important example of the utility of this dynamic flexibility, but one which helps to solve the growing issue for interconnection of data centers of balancing performance under burst demands with the need to avoid an expensive overbuild of the WAN.

Target System/Interface	Current Standards Activity	Gap
OpenFlow switching element (physical or virtual).	The OpenFlow protocol between SDN Controller and OF-enabled switching NEs for inspection of packets and setting of flow table entries is defined and being evolved by the Open Networking Forum (ONF). The PCEP protocol and Segment Routing are defined and being extended by the IETF for flexible traffic steering and service chaining.	None.
Network APIs to Multilayer SDN Control for orchestration of express traffic.	SDN Northbound APIs to “network controlling” applications are under consideration in various venues in the industry, including a new NBI-WG in ONF, but lag the development of Southbound (network-facing) interface specifications.	Expected to remain a near term gap, pending the successful completion of ONF NBI-WG work; not clear that additional parallel SDOs working NBIs would be productive at this point.

8 Conclusions & Recommendations

This focus group report has taken a “use case” approach in analyzing how network intelligence can be better leveraged in the network. Eleven different use cases have been analyzed. For each use case, both the network data inputs as well as potential outputs/network actions were considered to better leverage network intelligence. Input data was categorized as:

- Subscriber data;
- Application-specific data; and
- Network data.

Output/actions were categorized as:

- Output to user/application;
- Output to operations; and
- Network action.

Standards gaps were identified in the following areas:

Subscriber Data	Application Specific Data	Network Data
Subscriber data inputs identified in the Section 6 use cases where found to be generally addressed with existing standards associated with AAA, HSS, HLR and DHCP network elements.	• Identified need for a standard out-of-band mechanism for reporting application specific metrics from a device to either the network or application provider. • Additional standards work is required to create a robust framework for common application type definitions.	• Many IETF standards exist to aid in identifying the path of a flow through a network. However, this path information is generally not available across different routing domains. • Additional standards may be needed to enable the correlation of network element/link congestion data with subscriber identity. • No standard definition of RAN congestion in 3GPP. • Potential need for standardized collection and correlation of congestion data. • SIP Error Codes need to be further defined to accurately identify a blocked call condition. • Registration failures, ASR, and NER statistical data need to be further defined, filtered, and correlated to accurately identify OOS conditions.

Output to user/Application	Output to Operations	Network Action
Outputs/actions to users/applications were found to be generally addressed with existing mechanisms and standards.	No explicit output gaps identified, however; there is a need to align across the industry on best practices for analysis of network events to identify both out-of-service and blocked call conditions.	Open APIs to allow multi-vendor programmability by centralized network controlling applications to set policies for CAC and allocate RAN resources at the base station between different service provider or user group “slices”.

In the area of outputs and network actions, some near term gaps were identified that are believed to be resolved by ongoing NFV and SDN related standards.

Three common themes emerged as a result of the analysis. First, the correlation of input data can greatly improve the range and effectiveness of subsequent actions. For example, a node, link or device may report congestion. If this congestion event can be correlated to identify specific subscribers participating in the congestion and if the subscriber contribution to congestion based on what application is being used can be further correlated, there is a much richer set of potential mitigations that can better optimize network resources as well as create new monetization opportunities.

The second theme is that the emerging NFV and SDN work creates new opportunities to utilize network information by enabling the implementation automated network mitigations to “automatically” adjust network configuration and parameters to deal with changing traffic patterns and call models.

Finally, the Focus Group concluded that a common data exposure framework could increase the availability and usability of network data as collected by various analytics systems. That is, it would be useful if network elements could natively expose the right data in a common and consistent manner to allow analytics systems to capture the data with standard API calls within the network element. The focus group also realizes that such an undertaking may be a significant industry effort, however; this may be an

opportune time to address this as network elements are adapted for more programmable control. In any case, consistent and timely access to data is essential to analytics analysis.