Journal of Network and Systems Management (2020) 28:721–750
https://doi.org/10.1007/s10922-020-09517-0

Network Management 2030: Operations and Control of Network 2030 Services

Alexander Clemm[1] · Mohamed Faten Zhani[2] · Raouf Boutaba[3]

Received: 7 December 2019 / Revised: 17 February 2020 / Accepted: 21 February 2020 /
Published online: 3 March 2020

Abstract

The networking landscape is expected to undergo profound changes over the course of the next decade. New network services are expected to emerge that will enable new applications such as the Tactile Internet, Holographic-Type Communications, or Tele-Driving. Many of these services will be characterized by very high degrees of precision with which end-to-end service levels must be supported. This will have profound implications on the management of those networks and services, from the need to support new methods for assurance of ultra-high-precision services to the need for new network programming models that will allow the industry to move beyond DevOps and SDN towards User-Defined Networking. This article analyzes those implications and provides an overview of challenges along with possible solution approaches and opportunities for research.

Keywords High precision networking · Intent · Service assurance · Network operations · Service management · Network programming models · BPP · New IP · Research challenges

✉ Alexander Clemm
alex@futurewei.com; ludwig@clemm.org

Mohamed Faten Zhani
mfzhani@etsmtl.ca; Mohamed-Faten.Zhani@etsmtl.ca

Raouf Boutaba
rboubata@uwaterloo.ca

[1] Futurewei Technologies, Inc., 2330 Central Expressway, Santa Clara, CA 95050, USA

[2] Department of Software and IT Engineering, Ecole de Technology Supérieure de Montreal, 1100 Rue Notre-Dame Ouest, Montréal, QC H3C 1K3, Canada

[3] David R. Sheraton School of Computer Science, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada

1 Introduction

The networking industry is currently at an inflection point. 5G is being rolled out, leading to unprecedented bandwidth and ultra-reliable low-latency communications at the mobile edge that will enable many new applications. The Internet of Things (IoT) is exploding, transitioning our environment in cities, offices, and living spaces from static islands that are filled with passive objects into smart environments where everything becomes smart and interconnected. Artificial Intelligence and Machine Learning are finding real-world applications in the networking domain, promising not only greater operational efficiencies but improving networking services, for example, making communications more secure by automatically identifying and isolating threats.

At the same time, new challenges abound. Traditional business models and ecosystems of the networking industry are being threatened: providers of over-the-top services as well as new unregulated entrants (such as Amazon, Facebook, and Google) are challenging established service providers by providing offerings such as software-as-a-service that increasingly obviate the need for IT and VPNs, let alone voice and video services, while carrying traffic almost entirely across private networks of global scale without needing to rely on service provider infrastructure except perhaps (at least for now) for local access. At the same time, networking hardware is being commoditized by network virtualization and software-defined networking, creating challenges for equipment providers. The traditional Internet is increasingly marginalized by the emergence of "Manynets", i.e. private networks that operate within their own domain and in many cases no longer require global interoperability [1]. On the technical side, new networking applications such as Industrial Control, the Tactile Internet, or Tele-Driving promise new opportunities but are bumping into technical and physical challenges that are proving difficult to overcome, even with 5G, and that will require new solutions. These challenges are expected to have a profound impact not just on networking technology, but also on the management of networks and services.

For example, one of the next frontiers concerns high-precision networks that are able to provide very precise service level guarantees for end-to-end latency, such as needed for certain industrial controllers. Managing such services will, among other things, require advances in measurement technology to ensure very high accuracy with a very high degree of measurement coverage. Likewise, the trend towards making networks more programmable is expected to continue, ultimately allowing end users and applications to customize behavior for individual data streams and flows. Again, multiple ramifications for management can be expected, from the ability to monitor the actual communications behavior for compliance with the behavior intended by users, to the need to provide extended troubleshooting capabilities for AI-accelerated applications.

This article aims to look ahead at the trends and challenges that will shape network management in 2030. Its goal is to raise awareness of the shifts that are currently occurring and where they may lead, and to point out ramifications for management technology as well as to highlight open problems and opportunities for

research. Of course, as has been observed by Niels Bohr and others in the past, "it's tough to make predictions, especially about the future" [2]. Clearly some of our predictions may turn out to be wrong while missing other future networking trends entirely. Nevertheless, we are confident that many of the issues pointed out in this paper will be relevant and we hope to contribute to an understanding of their management implications.

The remainder of this paper is structured as follows: Sect. 2 will outline what the landscape for networks in 2030 is expected to look like and what emerging networking trends can be seen. On this basis, management ramifications will be analyzed and requirements for network management in 2030 articulated in Sect. 3. The focus will lie specifically on those aspects that are "new", not on aspects that are already well known today, even if they need to continue to be addressed going forward. Section 4 looks at future directions for management technology and problems that need to be solved in order to address those ramifications and new requirements. Where applicable, opportunities for network management research are pointed out. Section 5 concludes the paper.

2 Networks in 2030

To set the stage, this section will illuminate the networking landscape that providers and users of networking services and technology may be facing in 2030. Of course, many if not most of the same networking services that we have today will continue to play an important role in the future. Those services will continue to need to be managed, with continued management advances aiming at making their management ever-more efficient and cost-effective than before. However, in addition there are new networking themes which can be expected to emerge in the near future, or which have already started to emerge but have not yet come to full fruition. These are the areas in which we can expect to enter uncharted territory and encounter new challenges that need addressing and that have not been addressed before.

We start by providing an overview of the networking landscape that we expect in 2030. Subsequently, we flesh out emerging and recurring themes from that landscape which point to the need for new management solutions.

A. Networking Landscape in 2030

ITU-T has recently chartered a Focus Group on Network 2030, FG-NET2030, which examines a rich set of networking use cases that can be expected to emerge over the next decade and which specifies new networking services and capabilities that are expected to emerge in order to be able to address those use cases [3]. This analysis provides a perfect starting point to analyze the 2030 networking landscape.

1) New Services

FG-NET2030 expects a number of new networking services to emerge, which will in turn impose new requirements on management technology [4]. Predominantly this concerns services that have very precise timing requirements, both in terms of the end-to-end latency that can be incurred and the degree of synchronization required when multiple flows and data streams are involved. Early efforts that are just starting for the development of 6G [5] are also emphasizing this aspect, specifically the need for end-to-end ultra-reliable low latency communications (URLLC). While URLLC support has also been emphasized for 5G [6], that support focuses on the latency incurred at the edge and between end device and network (antenna and front-/mid-/backhaul), not between end-to-end communication peers and across the network core, which remains an open problem.

In addition, new services are foreseen that support refined structuring of payload and fine-grained prioritization of user data, including fine-grained control over selective dropping as well as retransmission of data when needed. This enables new schemes that minimize end-to-end latency by avoiding end-to-end retransmissions when not absolutely required, and that increase traffic resiliency by allowing to take better advantage of novel network coding schemes.

One type of service concerns Holographic-Type Communications (HTC), enabling networked applications that take advantage of advances in holographic display technology to build highly immersive networked applications. Far from being a gimmick, there are many useful scenarios for such applications. For example, holographic telepresence will allow a remote participant to be projected into a meeting room. Conversely, immersive spaces can project holograms resembling artefacts from distant locations to immerse the user into that space. Remote troubleshooting and repair applications could allow technicians to interact with holographic renderings of artefacts which are located in remote locations (e.g. an oil drilling platform, or an aircraft). Holographic signage used to render life-like signs presents a natural evolution for digital signage. Training and education applications can allow users to dynamically interact with ultra-realistic holographic objects for teaching purposes. Other applications may involve immersive gaming and entertainment.

Representations of holographic contents involve large volumes of data, resulting in large networking bandwidth demands. However, the precise contents that are visible to the user at any one point depends on the users position and angle, with a large portion hidden from view. This fact can be exploited to reduce and smartly compress holographic data that needs to be actually transmitted, as only those aspects of the hologram that are in focus for the user need to be rendered in high quality [7]. However, as position, viewpoint, and focus of the user may shift, it is important that the precise data that is being streamed can be adjusted very rapidly to maintain a high quality of experience. This requires very low end-to-end latency, not just for live contents involving interactions with another user, but even for prerecorded “canned” contents.

Haptic communications constitute another category of networking services, revolving around the transmission of tactile (involving a sense of touch, such as texture, vibration, and temperature) and kinesthetic (involving a sense of forces, such as gravitation and pull) data. An example of an application requiring telehaptics is telesurgery, in which a surgeon is able to perform a surgical procedure on a remote

patient. Arguably of much more massive relevance will be industrial applications in Industry 4.0 settings [8] that involve remote operations of machinery by a user. Haptic communications are particularly significant in the fact that they constitute a paradigm shift which takes networks beyond delivery of contents towards the delivery of skill sets and labor.

Haptic communications involves several channels. At its core, it includes a haptic feedback channel that is used to communicate haptic data streams from remote haptic sensors (for example, sensors in a robotic arm) to haptic effectors. This is typically complemented with a control channel used to operate remote actuators as well as with additional data feeds including visual (video, AR, holograms), acoustical, and telemetry. All of these streams need to be highly synchronized. In addition, tactile feedback involves stringent round-trip delays on the order of 5 ms or even less. Anything longer and the sense of remote touch and along with it the ability to confidently operate machinery from remote is lost.

Making these networking services a reality will involve advances not only in networking but also in management, with one of the biggest challenges concerning the assurance of service delivery according to service level objectives with much higher levels of precision than what was sufficient in the past.

2) New Infrastructure

Networked applications evolve in ways that involve ever more entities that are not only being interconnected, but that in many cases become an integral part of those applications themselves. For example, V2X communications interconnect vehicles as well as smart roadside infrastructure, resulting in smarter transportation and logistics services. The Internet of Things (IoT) continues its explosive growth, leading to integrated services from smart factories to smart cities, from agricultural optimization to remote home security. Applications that combine sensor and online data are beginning to mashup physical and virtual worlds. As things are getting "smarter", they are arguably becoming integral parts of the networked services and their underlying infrastructure themselves.

As more entities in the world are becoming interconnected and connectivity becomes an essential part of their function, connectivity is becoming increasingly ubiquitous as well. Satellite communications, drone communications [9], and ad-hoc vehicular networks can be expected to increasingly integrate with and complement mobile and fixed networks, patching any gaps in connectivity coverage. Likewise, satellite constellations that allow for traffic to be forwarded between satellites directly without the need to traverse ground stations between hops may become part of future infrastructure as well [10].

A wild card for new infrastructure concerns the possible emergence of quantum networking in which data between devices can be transmitted using quantum-entangled photons. This may eventually provide a solution for URLLC, particularly over longer distances, and obviate the need the need for traditional core routing infrastructure, which would in fact simplify management. At the same time, support for quantum networking will result in new types of devices with their own management

needs which are not yet fully understood. For example, the distribution of entangled photons may require separate infrastructure which will incur its own set of secondary management tasks [11, 12].

All of this implies that not only will the scale of networks and the number of networking devices continue to grow, but also that networking infrastructure is becoming far more diverse and with it the diversity of infrastructure that needs to be managed in integrated fashion. This is very different from the mere need to deal with device heterogeneity of the olden days, which to a large degree involved dealing with multiple variations of vendor interfaces and device capabilities. This aspect of heterogeneity has to a large extent been successfully addressed with the emergence of softwarized networks, which involved the introduction of virtualized and thus "standardized" infrastructure with homogenous interfaces, and with MSDCs (massively-scalable data centers), which involve largely standardized (within a given topology) configurations and topologies. In contrast, in the 2030 networking landscape, there will be a much wider gap in infrastructure component capabilities and management that needs to be bridged than was the case before, rooted in the greater diversity in the nature of the components that will need to interwork and that will collectively make up the infrastructure needed to provide novel networking services.

3) New Verticals and Business Models

Each new networking generation enables new business models and value chains, in turn resulting in new requirements and integration needs. For example, the rise of the World Wide Web enabled the Web economy and new companies such as Google and Facebook, with new business models centering around online advertising based on search, respectively social networking. Online shopping (Amazon) and video streaming (Netflix), enabled by broader availability of high networking speeds, are other examples. More recently, the emergence of mobile LTE (Long-Term Evolution) networks enabled the rise of Uber and Instagram and associated business models of sharing economy and social networks dependent on mobile apps. As new business models arose, so did new requirements and demands on infrastructure: business models relying on streaming of contents in turn drove advances in Content Delivery Networks (CDN), just like business models like Uber's lead to new requirements regarding the ability to integrate management of online and real-world (fleet management) infrastructure as well as advancing V2X communications.

By the same token, it should be expected that new networking services in a 2030 landscape will enable new business models that will in turn drive further infrastructure and management requirements. The precise nature of these will be hard to predict, but the fact that they are likely to emerge and result in new requirements needs to be acknowledged. As mentioned earlier, haptic communications services are a potential game changer in that they enable the delivery of skill sets and labor from remote. This may perhaps result in new business models related to providing medical procedures in rural areas where they were not available before. Coupled with audio-visual and telemetry feeds, haptic communications services may also enable teledriving services to complement autonomous

vehicles or services in which heavy machinery construction sites is operated from remote to increase construction safety. Holographic communications services might result in new business models such as Signage-as-a-Service or new types of remote collaboration services that span physical and cyber world. Also decentralized and federated AI and machine learning platforms and services are conceivable to let multiple interconnected nodes learn concurrently and from each other in real time, sharing neural network state across longer distances [13]. One use case is in future wireless networks, where edges collaboratively train a shared learning model using their respective data, without sharing that data for privacy concerns or limited resource constraints.

As a result, the scope and functionality of what integrated management needs to cover in order to provide holistic services as part of new business models is expected to grow further.

4) Towards Manynets

There are many other trends that will impact the 2030 networking landscape. In the past, one of the main driving forces for advances in the networking field has been network convergence: as the Internet grew into the dominant global network, and increasing number of services previously associated with their own separate infrastructures migrated onto it. Examples include voice (displacing Public Switched Telephone Networks—PSTN), video (streaming replacing traditional analog cable distribution networks), even SCADA (replacing traditional Supervisory Control And Data Acquisition for electrical grids with IP-based Smart Grid). This convergence was in large part rooted in network management and operational efficiencies that could be unleashed in needing to manage only a single, converged set of infrastructure. However, the drive towards convergence may not necessarily continue.

In recent years, massive global and private networks by Google, Amazon, and others have been emerging in parallel to the public Internet with end users to a large extent not communicating directly with each other but with services and servers hosted by those same providers themselves and using endpoints that are part of that same ecosystem. As a result, it is conceivable that those networks will evolve in separate ways, as within their own ecosystem they need to be less concerned with global interoperability. One example concerns the development of QUIC [14], a transport layer protocol originally designed by Google to provide a better user experience for interactions between the Chrome Web browser and Google's servers, replacing TCP.

Likewise, regulation increasingly results in the introduction of national boundaries into networks. Drivers include, for example, national data retention rules that require certain content to not traverse national boundaries, and different privacy rules and national regulations resulting not only in different operational policies but in separate sets of infrastructures with few and well-defined transition points.

As a result, the global Internet may give way to a "splinternet" [15] respectively to a set of "manynets". The Internet will continue to exist but be marginalized in

that it becomes only one of many networks, or just another service. By the same token, network convergence may no longer be a given. This will have ramifications also on the evolution of management technology, which may once again have to support diverging network infrastructures. As a result, it is conceivable that diverging management tools and interfaces themselves may potentially also begin to emerge.

B. Emerging and Recurring Networking Themes

There are several themes that emerge from the 2030 networking landscape, summarized in the following.

1) From Best-Effort to High-Precision

Many new networking services are characterized by stringent service level objectives (SLOs). In the past, deteriorating service levels generally resulted in gradual, graceful degradation of the associated Quality of Experience for end users. For example, for a streaming video service, video quality might be gradually reduced with deteriorating network service levels, resulting in lower resolution or color depth video quality. Web pages might take slightly longer to load. The quality of voice calls might be rated lower by users and result in lower mean opinion (MoS) scores. However, as long as service levels deteriorate only slightly, the underlying services are fundamentally still usable.

In contrast, many of the new services require strict adherence to service levels to be feasible. If service levels deteriorate, the ability to use the service rapidly breaks down completely, and with it the confidence of the user to even attempt using it. At the same time, many of the most interesting applications for those services are increasingly mission-critical and need to be highly reliable to be used at all. Teledriving, in which vehicles are operated from remote, and telesurgery are examples of such applications where even slight violations of SLOs quickly become matters of life and death.

This means that networks in 2030 need to move beyond best-effort services and support high precision: consistent delivery according to stringent service level guarantees, with hard boundaries regarding (for example) end-to-end latency and miss rates (i.e., rates of packets that are either lost or that violate their SLO) that are practically 0. Contrary to best-effort and QoS approaches of the past it will no longer be sufficient to merely "optimize" networks and service levels; instead quantified targets must be precisely met. By the same token, it becomes of utmost importance to assess in advance whether a network will be able to support the required guarantees, and to validate that guaranteed service levels are indeed being delivered.

None of this should be expected to come for free. Users demanding such services will be expected to pay a premium and provide incentives to ensure SLOs are adhered to throughout the lifetime of the service. Use for truly mission-critical applications may even give rise to demands to be able to insure such services, as they may expose providers to potential liability risk should they fail to deliver. The

necessity to move from best-effort to high-precision networks may thus also have ripple effects on other areas, such as accounting solutions and the ability to provide proofs of service level delivery.

2) Decentralization and Federation

Throughout the past, networking architectures have alternated between mostly centralized (e.g. PSTN, SDN controller-based architectures) and decentralized (e.g. Internet routing, edge compute). For the most part, this has been an architectural choice, trading off between considerations such as the ease of administration, the ease with which functional upgrades could be rolled out, the relative cost of coordination overhead, and the cost and complexity of system software.

This situation is about to change, as many new network services will require not only high precision but also very low latency. Contrary to services in the past, end-to-end latency budgets will no longer be measured in the 100 s of milliseconds (as was the case even for interactive services like telepresence or voice), but be at least an order of magnitude lower, sometimes in the single-digit milliseconds range. At this point, physical distance becomes an important consideration, as the speed of physical signal propagation is bounded by the speed of light (300 km per ms, even less in many physical media including optical fiber). This means that many of the new services will be restricted to scenarios in which communicating systems are located within a limited geographical radius.

As a result, there is a strong incentive to move XaaS ("X as a service") services, contents, and compute that are accessed via the network as close to the edge respectively to the receiver as possible. This means that it is no longer simply a choice whether or not to design architectures in a distributed, decentralized, or federated manner. Instead, it becomes mandated by the necessity to avoid having to service requests over long distances that would make it impossible to achieve low latency objectives. Architectures that depend on functionality which has to be provided at a central location are effectively ruled out by necessity. By the same token, networking architectures and management need to support and facilitate moving contents and services proactively as close as possible to the places where they will be requested and/or consumed. To be most efficient, management functions themselves may in turn become less centralized and more federated.

3) Unprecedented Scale and Scope

Another theme concerns the fact that the border between what constitutes a "networking service" and a "networking application" becomes increasingly blurred and in some cases irrelevant. Smart spaces, IoT/E (the Internet of Things, and the Internet of Everything), V2X (vehicle-to-x: infrastructure, other vehicles) communications all involve infrastructure beyond the traditional network which can be considered an integral part of the service. Accordingly, the scope of what a communication service entails changes and extends beyond that of the traditional network. Just like "cloud services" increasingly consider no longer compute separately and isolated

from networking and storage, "network 2030 services" will increasingly involve other infrastructure as part of a more holistic technological perspective. Of course, with growing scope, also the number of components that will need to be managed and the resulting scale continue to explode.

4) Agility

Agility has been a key driver for networking advances over the past decade. The emergence of network softwarization not only allowed to replace network appliances with virtualized network functions able to run on commodity hardware, resulting in considerable capital expense savings. Much more importantly, it opened the door for greater network development agility and enabled the emergence of DevOps, an integrated continuous development and network operations methodology. This empowered network operators to rapidly develop their own custom network adaptations and capabilities and introduce new services, breaking dependence on equipment vendor support and lengthy product development cycles for every new feature. As a result, the industry moved beyond "vendor-defined" towards "operator-defined" networking.

Going forward, the need for agility will only continue to grow, blurring the distinction between networking and management further. It should be expected that the need to customize existing networks or to introduce new networking behavior and communications service features will eventually extend beyond vendors and operators towards users and applications of networking services, moving the industry beyond "operator-defined" towards "user-defined" networking (Fig. 1). (In this context, "users" refers to customers of providers of future networking services, such as IT departments and providers of over-the-top (OTT) applications, not actual end users that simply consume communication services.) User-defined networking may ultimately result in the ability for users and applications to customize the behavior and introduce new network functionality on a per-flow basis. Instead of requiring

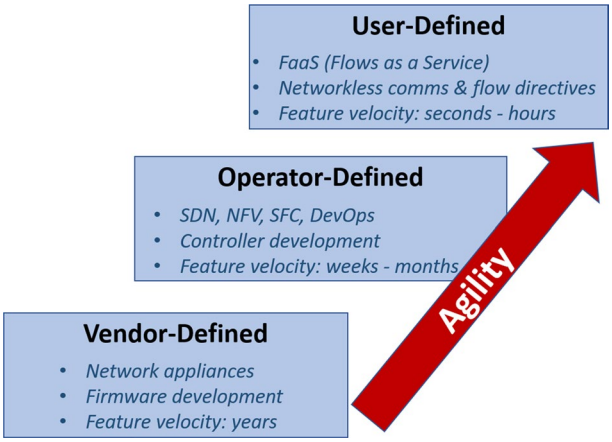

Fig. 1 From vendor-defined to user-defined networks

operator support for specific features needed for a new application or service, the operator provides merely the infrastructure over which end users and applications can run and adapt their own custom-defined communications services.

5) Privacy and Trust

Finally, with new services and applications making networking even more pervasive and increasingly fusing physical and digital worlds, concerns regarding privacy and the question of how to establish trust will become critical items of central importance much more than has been the case in the past. Strict new privacy rules such as the European Union's General Data Protection Regulation (GDPR) [16] are being introduced, while questions regarding digital privacy and trust are becoming top agenda items in global politics. Without convincing answers regarding how new services will address those concerns, the biggest challenge for new networking services may not be of technical but of societal nature.

As a result, privacy and the ability to establish trust between communication peers, and between users and the network, need to be addressed as an integral and inherent feature of new networking services. In general, new technology must be designed in ways that ensure privacy and security by design, as an inherent property of the service, not as an option or an afterthought. The success of new services will ultimately depend on it. Likewise, management itself is confronted with challenges. As inspection of user payload is off-limits and even communications meta-information such as present in packet headers is increasingly viewed as sensitive, corresponding data may not be available as input to management applications. Technology advances will be needed to bridge the legitimate need of network providers to understand what is happening in their network while ensuring privacy. This means they may potentially be prevented access to any data that could be processed in ways which might compromise privacy or trust.

3 Management Ramifications and Requirements

The newly emerging networking themes will have important longer-term ramifications on management technology. In addition, due to the nature of new networking applications and services, many new management requirements are introduced, some of which have never been seen before. In many cases, success and even the feasibility of those new services will depend on the ability of management technology to successfully address those challenges.

This section provides an overview of important ramifications and implications that the networking landscape in 2030 will bring. We will focus on newly emerging themes, not on existing management requirements and challenges which will continue to exist but which are already well understood.

A. Service Assurance and Visibility

One of the dominant and most striking network 2030 themes concerns the stringent service levels and high-precision assurances that must be supported. Advances in networking technology may be able to account to some degree for those requirements by moving beyond best-effort network principles and making high precision an intrinsic part of the technology. However, just as important will be advances in network management that will allow to manage networks and services in accordance with very tight service level guarantees. Assurance of those services will be of critical importance. This requires operators and assurance tools to have continuous visibility into service levels that are being delivered and to understand what is occurring in the network that may affect those services. Two areas of critical importance in that regard are measurements as well as telemetry data and visibility into flows.

1) Measurements

The ability to assure high-precision service levels depends on the ability to measure service levels with high precision as a prerequisite. Service level measurements are accordingly one area that will require advances. This concerns two aspects: precision and accuracy of the measurements themselves, and coverage of those measurements.

Today's measurements commonly rely on active measurements, in which test probes generate synthetic test traffic [17, 18]. This has the advantage of giving network providers full control over what to measure at what time, and allows service levels to be assessed proactively, including in advance of actual production (user) traffic. More importantly, it avoids the need to observe production traffic itself. This can be important to ensure compliance with privacy regulations, which may prevent network providers from "snooping" user traffic, let alone diverting copies of user traffic for later analysis. It also avoids needing to deal with implications of traffic encryption or tunneling, which may make it more difficult to discern between individual flows. However, one problem with active measurements concerns the overhead that is associated with synthetic test traffic, which consumes considerable resources in sending and receiving probes as well as traffic reflectors, in addition to consuming considerable network bandwidth. For this reason, measurements can only cover a sample of the network at any one point in time. Given the requirements for high precision, the possibility of misses due to statistical sampling may no longer be acceptable going forward. Instead, full coverage needs to be achieved without compromising measurement accuracy.

Passive measurements that rely on observations of production traffic, or hybrid schemes (in which production traffic is augmented with metadata used for measurement purposes), provide an obvious alternative. While coverage may be easier to accomplish here, there are other challenges: For one, the mentioned regulatory requirements may stand in the way of universal solutions. Likewise, achieving needed measurement accuracies (e.g. accuracies of 1 ms or less for end-to-end

latency measurements, and even less for measurements within individual devices) while avoiding performance hits on that production traffic can become challenging.

Hybrid measurement techniques have been proposed as well to combine the respective advantages of active and passive measurements while avoiding their drawbacks. Those techniques are typically based on marking production traffic with metadata used for measurement purposes, e.g. Packet Network Performance Monitoring (PNPM) and its variations [19]. However, the current state-of-the-art of those techniques is still limited and not well suited for many service levels beyond packet loss.

2) Telemetry and flow statistics

Another area that will require advances concerns support for telemetry and corresponding instrumentation. The ability to deliver high precision service levels for mission-critical services requires the ability to detect, understand, and counteract even slight fluctuations and degradations of service levels both at the flow and at the packet level. The same need simply did not apply for services in the past, for which statistical methods were sufficient and slight service level fluctuations were (within bounds) much more acceptable.

The problem to simply understand what is precisely happening to a given packet and how its service levels are being influenced is further compounded by the rise of virtualization. Processing of the packet may cross multiple virtualization boundaries. The length of Service Function Chains as well as associated networking paths may dynamically vary as VNF instances are migrated and SFCs are reconfigured. This introduces inadvertent variations in latency for different packets of the same flow, which is detrimental to achieving high precision.

Network telemetry and improved instrumentation have seen some important advances in recent years. For example, in situ OAM allows to collect critical performance measures from the network as a packet traverses a path [20]. YANG-Push allows to subscribe to continuous streams of network device data and statistics [21]. Distributed Network Analytics allows to dynamically adjust data to be generated at the source as needed and obtain more meaningful and actionable service assurance data [22]. At the same time, far more advances are needed. For example, there are limits to the frequency with which data snapshots of arbitrary size that are to be streamed from devices can be obtained using existing instrumentation technology. Likewise, the ability to collect comprehensive data on a per-packet level using iOAM techniques is still limited. Issues include impact on performance to retrieve the data, as well as the sheer potential volume of data (one data record for each packet and hop) coupled with severe limitations in the amount of data that can be piggybacked due to MTU considerations.

Another challenge for the collection of telemetry data concerns internetwork domain and trust boundaries. For example, while iOAM lets a packet collect data from the network, today this works only within a single network domain. Across multiple domains, no solutions exist. As trust boundaries are crossed, multiple concerns arise: how can a network provider expose telemetry data without also

exposing network internals to the outside? How can it be ensured that capabilities to collect telemetry data are not abused to attack a network, e.g. by generating heavy loads and creating new types of amplification attacks [23]? Can the data collected be trusted, or could it be falsely reported or forged while in transit? The latter becomes a concern specifically when high-precision guarantees for mission-critical services are involved, where any issues may result in significant legal and financial repercussions and where the incentive for tampering becomes high.

Similarly, solutions that provide telemetry data and statistics on a per-flow level are lacking. The state of the art in this area today is dominated by IPFIX and Netflow, which are very valuable tools but which suffer from important deficiencies. For one, they are computationally heavy. This means that practical deployments typically need to revert to sampling, not recording updates to flow statistics for every packet of a flow but only for a small sample, relying on statistical effects to result in flow data that is still "useful enough". However, the resulting margin of statistical error and imprecisions (with smaller flows possibly being missed entirely if their packets are not part of a sample) may be inacceptable for networking 2030 services. Another issue concerns the fact that the statistics provided are statically defined in advance and relatively coarse. In many cases the data recorded is simply not refined enough to meet the needs of precision services. Instead, facilities will be needed to customize and adapt flow and telemetry data dynamically and based on context in order to provide more actionable insights. Operational Flow Profiling [24] shows promise as one solution in this area.

It should be mentioned that all of those challenges are compounded by the fact that most network traffic will be carried across encrypted tunnels that may make it difficult to differentiate between individual flows, and not only payload (encrypted anyway) but also production traffic meta-information that is carried in headers is increasingly off-limits to network operators. One of the challenges will be to balance the legitimate need of network providers to understand what is going on in their networks and the service levels users are experiencing with the requirement for users to keep their communications private, including the fact that communications are even occurring.

B. Control and Fulfilment

Automated orchestration to fulfill services in real-time has arguably been the biggest enabler of networking advances in the recent past. Orchestration is a key enabler of SDN, allowing to provision flow tables, virtualized network functions, and service function chains in real time. Likewise, DevOps has revolutionized network operations by combining and integrating the continuous automation of operational work flows with agile development of extended networking features. As impressive as these achievements are, the emerging networking landscape of 2030 will require further advances.

Many of the new services that will need to be supported will be mission critical and require high precision. This means that it will become increasingly important

to not just configure the service itself, but to ensure that the required service level guarantees can be given as an inherent part of the fulfillment process. This includes steps to validate that service levels can indeed be attained as well as setting up facilities for service assurance, for example for continuous measurements. Rather than leaving these steps to separate service assurance processes, it will become important to address them as an intrinsic aspect of the service itself.

Real-time control will also be faced with new challenges. To meet high-precision requirements, novel congestion control and resource allocation and reservation schemes will be needed to minimize the possibility and the possible impact of competing service demands and non-deterministic disruptions. In addition, admission control schemes may need to be rethought. For example, before admitting a new flow or providing a service level guarantee, an admission control function could assess the likelihood that any violations might occur and make servicing of a communications request dependent on that outcome.

Ultimately, control and fulfillment will need to go beyond the mere automation of workflows and enter into the realm of "intent" (Fig. 2). Intent (in the networking context) is defined as the ability to allow users to define management outcomes, as opposed to having to specify precise rules or algorithms that will lead to those outcomes [25]. This requires an Intent-Based System to possess the necessary intelligence to identify the required steps on its own. In simple cases, a simple mapping or translation step similar to what a policy-based system would perform may be enough. In some cases, the translation steps may themselves result in network policies. However, more advanced and sophisticated systems may be able to apply artificial intelligence techniques to identify courses of action, dynamically moderate in real-time competing demands from millions of service instances, and apply learning techniques to optimize outcomes over time.

C. New Management Functions

Some of the newly emerging networking themes also imply the need for new management functionality in addition to classical management functions.

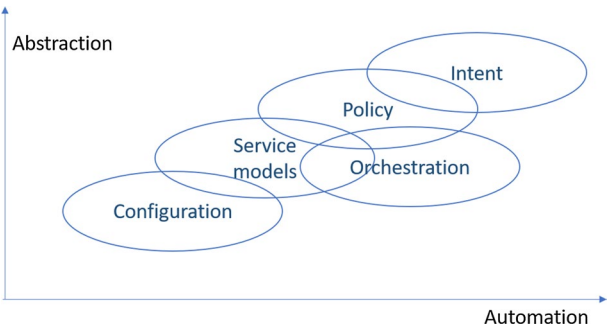

Fig. 2 Evolution of intent

For example, the growing emphasis on privacy will require ensuring privacy regulations are adhered to. Network providers will need to prove to users that they deserve the trust that their privacy is secured. Management functionality that will likely be required in the future will include assessment of compliance with privacy rules and best practices. Network analysis tools can be geared towards identifying points where traffic could be exposed to possible data leakage or snooping by other parts. Instrumentation will need to provide forensic data to validate what happened to network flows. Increasingly, proof-of-traversal functions will be required that allow to validate and prove which paths were traversed and which devices were touched, as well as which geographies and legal domains communication flows have been exposed to.

By the same token, due to the mission-criticality of many applications relying on high precision network services, validation that services were delivered in accordance with service level objectives will be increasingly required, along with proof of service levels for flows in any given network domain. Today's situation will simply become unacceptable, in which isolation of causes of service level degradation and violations is difficult (was the culprit the access network? the data center? the WAN link? the client or the server connected to the network?) and results in inconclusive finger pointing between different organizations involved in the service delivery chain. Instead, proof of service levels that are being delivered as part of every flow will need to be intrinsically addressed as part of providing such services, requiring support by corresponding management functions.

D. Management for Scale

The ability to manage networks at large scale has always been a challenge. This will remain true and become even more critical in the future. This is particularly compounded by two factors: for one, the scope of what entities must be included as part of a holistic management approach continues to grow and which increasingly extends beyond networking devices in a narrower sense and must include other

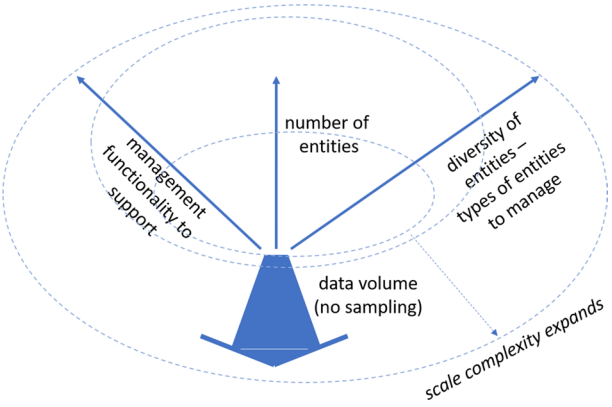

Fig. 3 Growing complexity of scale

artifacts, such as "things" (e.g. IoT) or traffic infrastructure (e.g. V2X). This does not only affect the number of entities needing to be managed and the resulting volume of data that needs to be dealt with, but also adds complexities due to their different nature. For example, the number of potential security threats and attack vectors on networks and their services is also exploding. Second, in the age of high precision services, service assurance techniques of the past that relied on sampling in order to scale will become increasingly insufficient as statistical gaps in coverage become less acceptable. This, in turn, results in even greater explosion of the volume of data (Fig. 3).

The implication of this is that technologies that facilitate management at scale will continue to rise in importance. This includes, for example, analytics and machine learning, which are instrumental at distilling seas of raw data into actionable information and directing the focus of management attention to where it is needed the most. This principle is proving increasingly indispensable in a wide range of data-driven management applications, from the automatic detection of traffic patterns that are out of the ordinary and might be indicative of malicious traffic or security attacks, to the optimization of resource allocation and network path computation in anticipation of user demand. The need for advances in this area will continue to grow. In addition, new challenges will need to be addressed. This includes the requirement to perform these functions under increasing privacy constraints, which may limit the data that is available to feed those functions, which may in turn obfuscate important information and limit certain conclusions that could otherwise be derived.

The need for scalability will also increasingly require management functions to not be performed from a single central location but to be distributed and possibly decentralized, increasingly pushed into the network and towards the network edge. Edge computing and fog computing have popularized the concept of moving processing to the network edge, close to users or sources of data being processed. In that sense, management can be considered as just another class of applications that can be subjected to the same principle. One such example is the before-mentioned Distributed Network Analytics [22], which pioneered moving network analytics to the source of network telemetry data, dramatically reducing data collection overhead while improving the quality of analytics in the process by custom-tailoring data sources dynamically depending on network context. Many more such examples will surely follow. Other drivers for decentralization include the need for management functions to be provided in increasingly shorter time scales, in particular in the case of high-precision services whose management may require closed control loops that are able to react in real time to, for example, realign resources or perform admission control functions. In conjunction with machine learning and analytics, this implies that techniques such as transfer learning (in which learned models from one context can be easily shared and adapted for other contexts) and federated analytics (in which data is selectively shared to reach common conclusions) will become increasingly important.

The need to scale extends beyond technology to network operations organizations. To support those organizations, management needs to continue to strive for simplicity and allow for management automation. Likewise, networks need to

continue becoming increasingly autonomic, eliminating the need for external systems or human intervention wherever possible. Management automation is an important topic in its own right and further discussed in Sect. 3(G).

Related to this is the topic of intent, mentioned already previously in the context of control and fulfillment (Sect. 3(B)). As networks become more autonomic, the importance of intent will continue to grow as network operators will still need to be able provide management guidance. Intent will let them focus on desired outcomes, as opposed to detailed instructions of what specific steps to take or which policies to follow. This in turn may very well turn out to be crucial in allowing network operators to keep up with the exploding scale of their task.

E. Accounting and Accountability

The rise of high-precision networking services used for mission-critical applications has important ramifications for another area of management, namely accounting. Providing service level guarantees and delivery of services according to high precision service level objectives cannot be expected for free; providers of those services will expect to be able to monetize them. Likewise, there needs to be accountability in cases when services are not delivered with the required service levels. This implies that delivery of communication services according to service levels objectives needs to be accounted for. Existing techniques based on the collection of interface statistics and flow records to determine volume of traffic will be no longer sufficient for those purposes. Instead, service level objectives need to be validated and adherence of network traffic to those objectives as well as any violations must be properly recorded as part of accounting data.

F. Programmability and Novel Programming Models

The demands for ever-increasing agility and the ability to customize network behavior and control down to the user and flow level that will be required to move beyond provider-defined to user-defined networking cannot be easily addressed with existing technology. In order to enable user-defined networking, major steps need to be taken towards further management simplification of the underlying networking infrastructure. (As explained in Sect. 2(B.4)), "users" refers to customers such as IT departments and providers of over-the-top application services, not actual end users that simply consume services). Users must be able to customize control of flows and fulfillment logic of specific service instances to allow for very specific outcomes. They also must be able to do so in a very simple yet secure manner that does not compromise underlying infrastructure or other users. This will require new, yet-to-be-developed network programming models.

Of course, to enable this, multiple advances must be made as a prerequisite. This includes ways to prevent abuse by users that would affect the network and expose its infrastructure, or that would compromise services provided to other users. In addition, this will also include ways that facilitate the accounting for services and networking resources that are consumed. Very importantly, it will

also require advances in network programming models that empower users to request custom networking behavior in ways that are extremely simple to define, observe, and understand.

In the adjacent field of cloud computing, one of the most recent advances concerns the emergence of functional programming and lambda functions ("serverless compute") [26]. This promises to be a game changer because it frees users from concerns about how to manage their virtual compute infrastructure. Up to this point, users had to dimension their virtual resources, orchestrate the spinning up of virtual machines and containers, provision their interconnections, and worry about the cost of resources being claimed but not utilized. Instead, lambda allows users to merely provide the compute functions that are to be provided, and the data to apply those functions to. Management of the underlying virtual infrastructure is no longer required; it is taken care of by the service itself. This is enabled by massive advances in autonomic management of the underlying infrastructure, which assumes all responsibility for proper configuration, dimensioning, allocation, and release (when no longer needed) of underlying resources.

The field of networking software has yet to see an equivalent for such services, but it will likely need to emerge as part of the 2030 networking landscape. While Network Function Virtualization (NFV), Virtual Network Functions (VNFs), Service Function Chaining (SFC), and Software Defined Networking (SDN) are powerful concepts, they come with their own set of complexities. These complexities should be hidden from users who may want to customize network behavior for their flows, but who should not have to worry about required configurations and who should be relied on to mitigate between their requirements and those of other services and users.

G. Automated Management

Automated management has been the holy grail of network management research for decades with the aim of closing the management loop and achieving autonomous networking, i.e., networks capable to autonomously monitor their status, analyze potential problems, make control decisions, and execute corrective actions. There have been several attempts to achieve self-managing networks, including policy-based management [27], autonomic networking [28], knowledge-driven networks [29], and recently self-driving networks [30]. However, practical deployments have largely remained unrealized. Several limiting factors can be attributed to this, including the existence of many stakeholders with conflicting goals, reliance on proprietary hardware and a complex web of interacting protocols, lack of global visibility restricting network-wide optimizations, and the inability to process network telemetry at scale.

The stars are now aligned to realize the vision of autonomous networking thanks to advances in network softwarization, recent breakthroughs in machine learning, and the availability of cloud platforms for large-scale data processing. However, these individual technologies are merely pieces of a bigger puzzle yet to be solved for the successful realization of autonomous networks. A number of

challenging issues need to be addressed not only to create the synergy between these different technology domains but also to develop a fundamentally new approach for the orchestration and management of softwarized networks. These include the ability to program the data plane in a protocol-independent manner for adaptive monitoring and control policy enforcement, real-time processing of streaming monitoring data, predictive machine learning for closed-loop network management, orchestration algorithms for cost-effective, resilient, and efficient service provisioning.

4 Future Directions for Network Management 2030 Technology

As discussed in the previous sections, the networking landscape in 2030 and associated newly emerging networking themes are expected to have signification management implications. This section highlights some of the future directions for management technology that will potentially play a prominent role in addressing future management challenges and which promise a rich set of opportunities for research and innovation.

A. Privacy-Preserving Management

Analytics and machine learning are becoming increasingly important tools that network operators depend on and that users benefit from, from the defense against security threats to the optimization of network services. At the same time, respecting privacy of user communications is becoming increasingly critical, extending not just to keeping payload private through encryption schemes, but even meta-information regarding what communications are happening in the first place. As a result, the needs of management tools that depend on visibility into network traffic and telemetry and the demands for privacy are seemingly diametrically opposed. Data available for management may be restricted to data that does not expose or allow to infer personal identifiable information. This provides challenges for a wide range of management functions, from the measurement of service levels to the detection of malicious traffic.

The challenge, then, concerns development of management technology that balances those concerns, ensuring privacy while at the same time letting network providers address legitimate operational concerns.

One technology that points into a promising direction is homomorphic encryption [31]. Homomorphic encryption allows certain operations to be performed on encrypted data, with the result of the operation guaranteed to be the same as if the encrypted data had been decrypted, the operation performed, and the result re-encrypted. To the best of our knowledge, there are currently no encryption schemes that allow management operations be conducted on encrypted data and achieve the same outcomes as if they were to operate on the actual decrypted data. However, the

development of such a scheme, even if custom tailored for only very specific management functions, would provide a giant technology leap forward.

B. Intent-Based Networking

In the last couple of years, the concept of Intent-Based Networking has emerged to extend the community's endeavor on policy-based management and autonomic networking. An *intent* refers to a high-level outcome (for example, an operational goal) set by the network operator and that the network itself needs to meet [25].

The concept of automatically breaking down management requests from higher levels of abstraction into low-level management actions has been applied by other technologies in the past, such as service order provisioning systems that break down requests for user services, or policy-based management, which allow operators to specify policies, often conditioned around rules that express what set of actions to take under which circumstances (often a combination of conditions and event triggers). However, in each case, the rules to apply or the mapping steps to take need to still be specified by a network administrator. In contrast, intent is about letting users specify desired outcomes, **without** having to specify the specific set of steps to get there or spelling out which actions to take under which condition. The set of actions to take or even the set of policies or algorithms to apply in order to achieve the outcomes may not be predetermined, but could be learned automatically by an Intent-Based (management) System over time. Likewise, the specification of intent by users may follow unconventional interfaces, not necessarily based on a traditional command syntax or request pattern, but allowing for human–machine dialog that allows for iterative refinement and includes explanation components. These aspects set Intent-Based Networking apart from other technologies before it.

While the concept of Intent Based Networking seems to be appealing, many research challenges still need to be addressed before it can see the light. These challenges include the management of the intent lifecycle, starting from defining the right interfaces to describe intent, to assessing and validating whether the network is

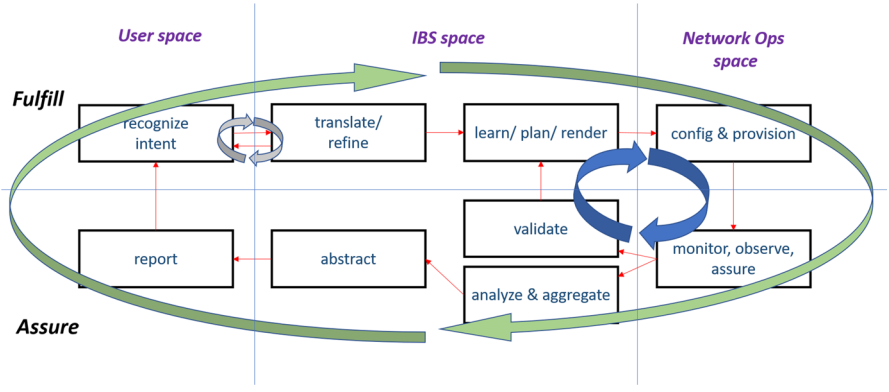

Fig. 4 Intent lifecycle (per [25])

indeed complying with intent, to intent rendering and maintenance. An example of an Intent Lifecycle adopted from [25] is depicted in Fig. 4.

The aspect of intent interfaces raises many interesting research challenges. These include question of how to let the Intent-Based System to interact with users in order to let them refine their intents, to better clarify their needs, and to inform them of the possible ramifications of the implanted intents. In other words, Intent interfaces should provide tools enabling the users to interact and negotiate to support them in defining what they really aim at. In other words, instead of simply executing what the user says, the goal is to achieve what the user actually wants.

Another key challenge pertaining to intents is how to let the management system automatically render an intent, i.e., how to translate an intent into low–level network configuration, rules and actions. Addressing this challenge calls for the design of autonomic management systems able to map the desired Intent outcome into device-specific instructions and to identify actions to coordinate between the involved nodes in the network in order to achieve the intent's outcome. In this context, Artificial Intelligence and Machine Learning techniques could be a valuable tool to automatically learn and refine smart algorithms to adjust network configurations and take the right course of actions to continuously achieve the sought–after outcome of the intent.

Furthermore, research efforts should also explore different intent rendering solutions. For instance, they can investigate using centralized solutions where a single component is in charge of rendering the intent compared to a distributed approach where several components render the intent in different parts of the network and cooperate in order to implement the desired intent outcome. Another interesting avenue to explore is intent rendering solutions that take into consideration the heterogeneity of the underlying network infrastructure. This is particularly challenging as network equipment are heterogeneous in terms of technologies, interfaces and performance and may have different capabilities that should be taken into account in the rendering process.

Networks in 2030 are likely to continue to rely on network softwarization, with technologies like SDN and NFV continuing to play a role. The deployment of these technologies stipulate that network services will be offered as service function chains that are composed of different types of virtual network functions running as software as software appliances in virtual machines or containers. As a result, translating intents into the appropriate service function chains as well as identifying, configuring, provisioning and chaining their constituent virtual network functions is another key research problem to be addressed in order to realize Intent-Based Networking in softwarized networks.

Once an intent is implemented, it requires maintenance. Network and the service performance must be monitored continuously to ensure that the service requirements are satisfied and service level objectives are met. This may require the management system to dynamically adjust resource allocations and network configuration. It is of the utmost importance to develop systems that monitor intent implementation to ensure its expected outcomes are achieved.

C. Advances in Accounting Management and Incentive-Based Service Delivery

Accounting management aims at tracking the usage of services and the amount of used resources and charge users based on some pricing scheme. This is particularly challenging in future networks where service function chains are deployed dynamically using various types of resources including bandwidth, CPU, memory, storage, GPU, and Neural Processing Unit (NPU). The diversity of resources poses challenges as to how efficiently track the usage of each of them and estimate their costs taking into account several parameters like the resource type and usage, energy consumption, and performance. This requires defining novel pricing models that take into consideration the whole service chain as well as the nature and costs of its constituent network functions, the type of used resources, their utilization and the overall requested performance.

Future networks should also implement efficient proof–of–delivery schemes to validate that provided service levels and performance satisfy users' expectations. In this context, an interesting research avenue is to investigate the deployment of escrow schemes where a third party monitors the service performance and ensures that the network operator pays back a penalty if the service level agreement is not satisfied.

Successful delivery of high precision services poses many technical challenges. Issues such as QoS assurance and congestion control based on allocation/reservation of resources, admission control, traffic engineering and provisioning of service chains have been subject to extensive research in the past. Quite possibly, those schemes by themselves will not be sufficient and alternative approaches will be needed. In addition, a problem with practical deployments is that many QoS schemes may not even be adequately followed, in particular when multiple domains are involved. For example, a common technique involves using IP's Differentiated Service Code Point (DSCP) field to mark a packet order to indicate that a packet should be handled in a certain way. However, devices may or may not always act accordingly. In cases where a packet crosses an organization boundary, any DSCP markings are often simply ignored. The underlying problem is that the interest of the client or border edge router (which wants a certain treatment for their packet) is in many cases not aligned with the interest of other devices and not consistent with that of domains that are crossed (which may have a different idea about whose traffic should be prioritized, for instance).

One interesting alternative approach concerns incentive-based schemes in which devices in the network are provided with "incentives" to deliver certain packets or certain flows in ways that ensure compliance with given objectives, for instance by prioritizing them or making smarter forwarding decisions that take the dynamic network context into account. Other incentive-based schemes might apply post-pay concepts that provide a "payout" in case of successful service delivery. In order to enable such schemes, support for service-level accounting data and proof-of-delivery (according to a service level guarantee) schemes will play a key role. In addition, to prevent abuse and combat fraud, it is conceivable that novel network-embedded escrow functions will emerge that provide trusted and tamper-proof validation and assignment of incentives upon successful service delivery. Such concepts are still

in their infancy and require more research before their feasibility can be positively affirmed.

D. Data Proliferation Management

As outlined in Sect. 2.B.5), concern for privacy is one of the themes whose relevance will continue to grow. This affects how management operations need to be conducted as it can no longer rely on data that could potentially compromise privacy, as explained in Sect. 4.A. Another implication concerns the need for management functions that provide visibility of what happens to networking traffic. For example, it may be required to be able to detect whether network traffic could cross geographical boundaries, or to identify points where data leakage might occur. While this is a largely unexplored area today, it is quickly increasing in relevance. Proof-of-transit technology [32] which allows to prove which network devices are being traversed by a given packet is one noteworthy development in that direction; more can be expected to come.

E. Novel Network Programming Models

Since the genesis of the Internet, network programming have been based on traditional socket programming that provides Application Programming Interfaces (APIs) for developers to use mainly, if not only, UDP and TCP over the IP protocol. This model relies also on several principles that have been always considered as the basis of network communications like the client and server model (i.e., a communication involves only two parties) and the end–to-end principle (application- and transport-layer features and data lie only at the end points). However, recent research work is calling into question such traditional principals and the efficiency of traditional protocols for future applications [33, 34]. Network 2030 are hence expected to incorporate a new generation of protocols and network programming models that are more adapted to the characteristics and requirements of the future applications and services (e.g., Haptics, Holoportation and high-precision communications) and that could leverage the recent technologies and trends (e.g., network softwarization, network function virtualization and in–network computing).

In this context, a novel communication protocol, Big Packet Protocol (BPP) [33], has been recently proposed as part of a larger networking framework, New IP. BPP allows programming network services from the network edge and to allow users to define and customize the network behavior per packet, flow or network service. This is carried out by incorporating into the packets metadata (i.e., additional information to be used by the network) and even commands that can be executed by the network nodes. While BPP offers an unprecedented flexibility and possibility to dynamically program, configure and adjust the network behavior and services, traditional socket programming does not provide the right APIs to easily implement and leverage BPP features. There is therefore a compelling need to develop sophisticated socket programming with additional APIs that could easily allow to implement additional

application-specific features and to inject their associated BPP commands and metadata into the communications.

Another recent related work is the Flexible Next–Generation Internet architecture (FlexNGIA) [35] which advocates to provide the network applications' developers with the ability to design and develop not only the software at the end points but also the network functions implemented within the network as well as the communication protocols (layer 3 and above). Consequently, FlexNGIA considers that communications could involve several end–points (not limited to two) and that the end-to-end principle is not necessarily respected as the network functions incorporated in the intermediate nodes could manipulate the upper layer data and could also implement any of the services traditionally offered by the upper layers of the OSI model. It is clear that these FlexNGIA features cannot be implemented using traditional socket programming as the basic principles of traditional network programming have been altered. Future research should therefore concentrate on the design of novel socket programming that can be fully customized to support any communication protocol at any of the protocol stack layer and that can inherently support multiple end–points and take also into consideration the existence of intermediate network functions.

F. Softwarization Interplay with Hardware Advances

Although the trend towards network softwarization has been gaining momentum in the last few years, the performance of network functions implemented as software appliances is still questionable compared to that of their hardware counterparts. There are legitimate concerns about the performance, reliability and scalability of software-based solutions. For instance, applications such as holoportation with stringent requirements in terms of throughput and latency may not be feasible with software-based network functions because of their limited throughput and high processing time.

One immediate research direction to address such limitations is to enhance the performance of software-based solutions in terms of throughput and processing delays. This requires a careful analysis of the characteristics and requirements of the various network functions and to re-engineer the full software stack including network function software, operating systems, hypervisors, IOs and network drivers leveraging new programming models and taking advantage of parallel processing techniques. More research work is also needed to define benchmarks and experimental methods to assess, model and predict the performance of software-based network functions especially when running on different hardware platforms with various technologies and capabilities. Few recent works started looking at the aforementioned challenges [36–38] but they constitute only a first step in this direction.

Another research direction to address the performance limitations of software-based solutions is to and implement network functions in a programmable data plane (e.g., [39]) leveraging programmable hardware technologies and platforms

such as NetFPGA and P4. This in turn requires new advances in hardware programming languages, data structures and algorithms.

One interesting case in point concerns BPP, the novel packet programming framework and protocol mentioned earlier. BPP allows to carry directives as part of packets in a flow to guide their processing, which can be used for a wide variety of purposes such as achieving high-precision latency-based forwarding [40] or for greatly improving operational visibility into flows [24]. Directives can be parametrized and subjected to conditions. This results in much more powerful functionality than can be accomplished with other programmable data plane technology. However, optimization the processing in hardware is a challenge as packet processing is not easily mapped into a conventional pipeline with serialized stages that have constant processing cycles. BPP allows for multiple directives, conditions, and parameters. While concurrency can be exploited to optimize processing, doing so in practice will require further hardware advances.

G. Flexible Network Monitoring

A network monitoring probe collects data to compute specific metrics, e.g., link utilization. It can be used as a building block for composing more complex monitoring queries. A probe can be deployed on both traditional and programmable switches that provide better visibility into the traffic, but are resource constrained or on end hosts that have more resources but reduced visibility. Therefore, network monitoring probes need to be intelligently distributed on end hosts and programmable switches for maximizing network visibility while operating under resource constraints (e.g., CPU, memory, flow tables). Another way of addressing the resource constraints of programmable network devices is to leverage streaming data structures such as sketches with bounded memory for approximate measurement. One challenge here is to make these sketches generic while ensuring a theoretically proven bounded accuracy.

The increasing ability to program the data plane will result in significantly more network monitoring probes deployed in the network and consequently a significantly increased monitoring overhead. This stresses the need for monitoring algorithms that have minimal footprint, without sacrificing accuracy. Combining sampling and ML-based inference of monitoring data for improving monitoring accuracy while reducing overhead is a promising research direction.

Networks of the 2030's must be equipped with the capability of automatically composing monitoring queries from high-level requirements of network management applications. For instance, a performance management policy specifying a bounded delay between a pair of network nodes translates into a query to probe specific links, queues, and other relevant delay parameters. Instead of human operators manually generating monitoring queries using domain specific languages, data models and languages are needed for capturing monitoring requirements of management applications and automatically generating monitoring queries from these high-level requirements.

H. Knowledge Extraction and Automated Decision Making

The value of monitoring data lies in the knowledge that can be extracted from it to predict network behavior, detect anomalies, and answer what if questions such as the impact of adding a new service. The challenge here is to identify and collect relevant monitoring data for different management functions such as failure detection (e.g., switch queues, alarms, event logs, and possibly other monitoring probes deployed in programmable data planes) and develop scalable ML-based solutions to detect and interpret anomalies in network behavior, e.g., network congestion, network partitioning, etc. Furthermore, these ML models will also need to adapt to changing network configurations and application mix without the need to retrain the models from scratch, which is still an open challenge.

Once anomalies such as failures, performance degradations and security threats are detected, it is essential to identify the root cause in a timely manner and deploy the appropriate mitigation plan to minimize impact. This requires an accurate representation of the network state (e.g., workload, enforced policies, configuration) at the time the failure occurred. However, this is challenging since hidden correlations within and between a large number of high-dimensional network state variables need to be uncovered. Existing ML-based root cause analysis approaches suffer from poor scalability [41]. In this area, promising approaches that can scale to large networks include those relying on cascaded Deep Learning models since they can be trained in parallel with less training data.

Once the root cause of a failure is identified, the next step is to automatically decide a mitigation workflow. Traditional approaches using "if-condition-then-action" rules designed by domain experts will be infeasible for the 2030's networks because it is far too complex to decide the optimal workflow of actions in every possible network condition. ML can be helpful in this context however even existing ML-based solutions cannot scale to the large state-action space of mitigation workflows [41]. Deep Reinforcement Learning methods are more promising in this case for their ability to handle the high-dimensionality of the state-action space. In any case, another challenge is the need for methods capable of handling unstructured operational logs in free-text format with possibly missing information.

I. Agile and Resilient Service Orchestration

Though current practice of replacing hardware middleboxes with monolithic VNFs is a good step forward to achieve better flexibility and maintainability as well as reduced capital and operational expenditures, it is far from being optimal [42]. Reliance on monolithic VNFs results in redundant development of packet processing tasks across VNFs and coarse-grained resource allocation. VNFs can be re-architected by disaggregating them into independently deployable packet processing entities following the microservices approach. Microservices structure an application as a collection of loosely coupled services and has been proven effective for building large cloud applications. VNFs and SFCs will then be realized by composing a packet processing pipeline from these independently deployable entities.

Ensuring network service reliability is another compelling issue. Purpose built hardware middleboxes have a proven track record for reliability; this is not the case for software VNFs running on commodity hardware. Service outages due to VNF failures result in significant financial loss [43]. Designing fault-tolerant network services is challenging because it requires fast and full recovery of VNF state after failures. An alternative approach is to replicate a VNF state at other VNFs along the same chain instead of using a per VNF dedicated replica. This approach would be a fundamental departure from the state-of-the-art for it considers SFC as the unit of fault tolerance and exploits the chain structure for state replication. It has the potential to achieve fast recovery after failures and significantly reduce latency during failure-free operations.

5 Conclusions

The networking landscape is expected to undergo significant changes over the coming decade. New types of services are expected to emerge, many of which will be defined by their need for high precision and used by mission-critical applications, which will push existing "best effort" network technologies to their limit. This and other emerging trends are expected to have profound implications also on management technology and on the way networks and services are managed.

In this article, we laid out many of those implications and their associated technical challenges. These challenges include but are not limited to the assurance of high-precision service levels, which will require advances in measurement technology and generation of network telemetry data, new management functions that are able to better address concerns about privacy, new network programming models to support greater network agility, and the requirement for novel approaches to operational scale such as Intent-Based Networking. Each of these challenges promises exciting opportunities for innovation. While some promising approaches have begun to appear, many unsolved problems remain. We hope that this article will provide a small contribution to this area by stimulating further much-needed research.

References

1. Huston, G.: The death of transit and beyond. https://www.enog.org/wp-content/uploads/presentations/enog-13/3-2017-05-25-death-of-transit.pdf (2017). Accessed 4 Dec 2019
2. https://quoteinvestigator.com/2013/10/20/no-predict/. Accessed 30 Jan 2019
3. Network 2030—a blueprint of technology, applications and markeet drivers towards the year 2030 and beyond. White Paper, https://extranet.itu.int/sites/itu-t/focusgroups/net-2030/SitePages/White%20Paper.aspx. ITU-T FG-NET-2030, (2019)
4. ITU-T FG-NET2030: New services and capabilities for network 2030: description, technical gap and performance target analysis. FG-NET2030 document NET2030-O-027, (2019)
5. O'Brien, C: Why 6G research is starting before we have 5G. https://venturebeat.com/2019/03/21/6g-research-starting-before-5g/). Venturebeat, (2019)

6. Li, Z., Shariatmadari, H., Singh, B., Uusitalo, M.: 5G URLLC: design challenges and system concepts. 2018 in 15th International Symposium on Wireless Communication Systems (ISWCS), IEEE, August 2018
7. Clemm, A., Torres Vega, M., Kumar Ravuri, H., Wauters, T., De Turck, F.: Towards truly immersive holographic-type communication: challenges and solutions. IEEE Commun Mag. **58**(1), 93–99 (2020)
8. Zhou, K., Liu, T., Zhou, L.: Industry 4.0: towards future industrial opportunities and challenges. in 12th International Conference on Fuzzy Systems and Knowledge Discovery (FKSD), IEEE, (2015)
9. Gharibi, M., Boutaba, R., Waslander, S.: Internet of drones. IEEE Access **4**, 1148–1162 (2016)
10. Guidotti, A., Vanelli-Coralli, A., Foggi, T., Colavolpe, G., Caus, M., Bas, J., Cioni, S., Modenini, A.: LTE-based satellite communications in LEO mega-constellations. Int. J. Satell. Commun. Netw. **37**(4), 316–330 (2019)
11. Why the quantum internet should be built in space. https://www.technologyreview.com/s/614994/why-the-quantum-internet-should-be-built-in-space/. MIT Technology Review, (2020)
12. Khatri, S., Brady, A., Desporte, R., Bart, M., Dowling, J.: Spooky action at a global distance—resource-rate analysis of a space-based entablement-distribution network for the quantum internet. arXiv preprint arXiv:1912.06678 [quant-ph], (2019)
13. Ready for 6G? How AI will shape the network of the future. https://www.technologyreview.com/s/613338/ready-for-6g-how-ai-will-shape-the-network-of-the-future/. MIT Technology Review, (2019)
14. Iyengar, J., Thomson, M.: QUIC: A UDP-based multiplexed and secure transport. IETF draft-ietf-quick-transport-25, (2020)
15. Malcomson, S.: Splinternet: How geopolitics and commerce are fragmenting the World Wide Web. ISBN 978-1-682190-30-2, OR Books, (2016)
16. General Data Protection Regulation. https://en.wikipedia.org/wiki/General_Data_Protection_Regulation. (2019)
17. Hedayat, K., Krzanowski, R., Morton, A., Yum, K., Babiarz, J.: A two-way active measurement protocol (TWAMP). IETF RFC 5357, (2008)
18. Chiba, M., Clemm, A., Medley, S., Salowey, J., Thomare, S., Yedavalli, E.: Cisco service-level assurance protocol. IETF RFC 6812, (2013)
19. Fioccola, G., Capello, A., Cociglio, M., Castaldelli, L., Chen, M., Zheng, L., Mirsky, G., Mizrahi, T.: Alternate-marking method for passive and hybrid performance monitoring. IETF RFC 8321, (2018)
20. Brockners, F., Bhandari, S., Pignataro, C., Gredler, H., Leddy, J., Youell, S., Mizrahi, T., Mozes, D., Lapukhov, P., Chang, R., Bernier, D., Lemon, J.: Data fields for In-situ OAM. Internet Draft draft-ietf-ippm-ioam-data-08, IETF, (2019)
21. Clemm, A., Voit, E.: Subscription to YANG notifications for datastore updates. IETF RFC 8641, (2019)
22. Clemm, A., Chandramouli, M., Krishnamurthy, S.: DNA: an SDN framework for distributed network analytics. in 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM), (2015)
23. Rossow, C.: Amplification hell: revisiting network protocols for DDoS abuse. in Network and Distributed System Security Symposium (NDSS) (2014)
24. Clemm, A., Chunduri, U.: Network-programmable operational flow profiling. IEEE Commun. Mag. **57**(7), 72–77 (2019)
25. Clemm, A., Ciavaglia, L., Granville, L., Tantsura, J.: Intent-based networking—concepts and overview. draft-clemm-nmrg-dist-intent-03, IETF, (2019)
26. Baldini, I., Castro, P., Chang, K., Cheng, P., Fink, S., Ishakian, V., Mitchell, N., Muthusamy, V., Rabbah, R., Slominski, A., Suter, P.: Serverless computing: current trends and open problems. in Research Advances in Cloud Computing, Springer, Singapore, (2017)
27. Boutaba, R., Aib, I.: Policy-based management: a historical perspective. J. Netw. Syst. Manag. **15**(4), 447–480 (2007)
28. Dobson, S., Denazis, S., Fernandez, A., Gaıti, D., Gelenbe, E., Massacci, et al.: A survey of autonomic communications. ACM Trans. Auton. Adapt. Syst. **1**(2), 223–259 (2006)
29. Clark, D., Partridge, C., Ramming, J.C., Wroclawski, J.T.: A knowledge plane for the internet. ACM SIGCOMM, Karlsruhe, Germany, (2003)
30. Juniper Networks: The self-driving network. White paper, (2017)

31. Lauter, K., Naehrig, M., Vaikuntanathan, V.: Can Homomorphic Encryption be Practical? in 3rd ACM Workshop on Cloud Computing Security (CCSW'11), Chicago, IL, (2011)
32. Brockners, F., Bhandari, S., Mizrahi, T., Dara, S., Youell, S.: Proof of Transit. draft-ietf-sfc-proof-of-transit-04, IETF, (2019)
33. Li, R., Clemm, A., Chunduri, U., Dong, L., Makhijani, K.: A new framework and protocol for future networking applications. in ACM SIGCOMM Workshop on Networking for Emerging Applications and Technologies (NEAT), Budapest, Hungary, (2018)
34. Tulumello, A., Belocchi, G., Bonola, M., Pontarelli, S., Bianchi, G.: Pushing services to the edge using a stateful programmable dataplane. in 2019 European Conference on Networks and Communications (EuCNC), Valencia, Spain, pp. 389–393. (2019)
35. Zhani, M.F., Elbakoury, H.: FlexNGIA: a flexible internet architecture for the next-generation tactile internet. arXiv 1905.07137, (2019)
36. Chowdhury, S.R., Bian, A.H., Bai, T., Boutaba, R.: μNF: A Disaggregated Packet Processing Architecture. in IEEE Conference on Network Softwarization (NetSoft 2019), Paris, France, (2019)
37. Ghrada, N., Zhani, M.F., Elkhatib, Y.: Price and performance of cloud–hosted virtual network functions: analysis and future challenges. in IEEE Performance Issues in Virtualized Environments and Software Defined Networking (PVE-SDN NetSoft 2018), Montreal, Canada, (2018)
38. Rosa, R., Rothenberg, C.E.: Automated VNF Testing with Gym: A Benchmarking Use Case. 1–2. https://doi.org/10.23919/tma.2018.8506566. (2018)
39. Sivaraman, A. Subramanian, S., Alizadeh, M., Chole, S., Chuang, S., Agrawal, A., Galakrishnan, H., Edsall, T., Katti, S., McKeown, N.: Programmable packet scheduling at line rate. in ACM SIGCOMM, Florianapolis, Brazil, (2016)
40. Clemm, A., Eckert, T: High-Precision Latency Forwarding over Packet-Programmable Networks. in IEEE/IFIP Network Operations and Management Symposium (NOMS), Budapest, Hungary, (2020) **(to appear)**
41. Boutaba, R., Salahuddin, M.A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F., Caicedo, O.M.: A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. J. Internet Serv. Appl. **9**, 16 (2018)
42. Chowdhury, S.R., Salahuddin, M.A., Limam, N., Boutaba, R.: Re-architecting NFV ecosystem with microservices: state-of-the-art and research challenges. IEEE Netw. **33**(3), 168–176 (2019)
43. Potharaju, R., Jain, N.: Demystifying the dark side of the middle: a field study of middlebox failures in datacenters. in ACM Internet Measurement Conference, Barcelona, Spain, (2013)

Alexander Clemm is a Distinguished Engineer at Futurewei in Santa Clara, California. Recent interests include high-precision networks, future networking services, network analytics, intent, and telemetry. He regularly serves in the OCs of IM/NOMS, CNSM, and NetSoft. He has 50+ patents, 12 RFCs, and a Ph.D. from University of Munich, Germany.

Mohamed Faten Zhani is Associate Professor of Software and IT Engineering at ÉTS Montreal (Canada). His research interests include cloud computing, network function virtualization, software-defined networking and resource management. He is co-editor of the IEEE Communications Magazine feature series on Network Softwarization and vice-chair of the IEEE Network Intelligence Initiative.

Raouf Boutaba is University Chair Professor in Waterloo (Canada) and INRIA International Chair (France). His research interests are in the areas of network management and resource virtualization. He is fellow of the IEEE, the Engineering Institute of Canada, the Canadian Academy of Engineering, and the Royal Society of Canada.