NSF Follow-on Workshop on Ultra-Low Latency Wireless Networks

November 3-4, 2016

Report from the National Science Foundation funded workshop held November 3-4, 2016 in Arlington, VA to address challenges in the design of ultra-low-latency wireless networks.

This material is based on work supported by the National Science Foundation under grant number 1448360. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Executive Summary

This report outlines findings from the National Science Foundation funded workshop on Ultra-Low-Latency Wireless Networks, held in Arlington, VA on November 3-4, 2016.

Wireless networks have become a ubiquitous part of everyday life all around the world. Yet, wireless networks are very unpredictable on one critical aspect: communication latency. It is well-known and widely observed that the latency incurred in accessing a wireless network can vary widely. The issue of unpredictable and often high latencies precludes wireless networks from being used in mission-critical environments.

Today's communication networks are largely geared towards latency tolerant (web, chat, email) content. Thus, these networks have been typically engineered with a focus on improving network capacity, with little attention to latency. However, in a range of domains, a wave of socially useful applications are emerging based on automated sensors and actuators operating in closed-loop or open-loop control systems. In these systems, including internet of things (IoT) applications, vehicular networks, smart grid, distributed robotics, and other cyber-physical systems, the requirements for latency could be two or three orders of magnitude more stringent than traditional applications. In addition, there are immersive services such as augmented reality that also require latency much smaller than what is achievable in today's wireless systems.

An earlier workshop in March of 2015 outlined the technical challenges and opportunities for achieving end-to-end low latency over wireless networks. This follow-on workshop was focused on emerging applications of: Telesurgery/Telemedicine, Manufacturing, and Augmented/Virtual Reality with the goal of understanding the critical requirements of these application domains, and developing feasible architectures and technological solutions for meeting these requirements.

Section 2 of the report addresses the application requirements of these emerging applications in terms of throughput and latency and reliability. With these requirements in place, Section 3 of the report addresses architectural alternatives for meeting these requirements. Finally, Section 4 of the report addresses technology limitations and opportunities for meeting these requirements.

Telesurgery and telemedicine encompass several applications ranging from remote diagnosis to remote consulting during a procedure to true remote surgery. On the patient side, this application may require several cameras (some of which could be wearable by the local staff). On the remote expert/surgeon side, these applications may require a VR interface in order to provide an immersive sense of experience. The latency requirements of telesurgery are outlined in Table 1, and range from 1ms to 100ms.

Augmented Reality (AR) adds supplemental elements (computer-generated) to the user's viewpoint, while Virtual Reality (VR) creates the entire scene the user sees based on multiple sensor sources (*e.g.*, cameras and other sensors). Both technologies operate in real time, and the latency requirements of these applications are outlined in Table 2 and range from sub millisecond to 10ms.

Manufacturing systems arise in such diverse domains as automotive/aerospace manufacturing, high-speed semiconductor manufacturing, smart-grid control, additive manufacturing, etc. Accordingly, the networking needs of these systems can range vastly in their scale and requirements. On the one hand, the scale can range from a single or a small group of co-located devices to a large plant which encapsulates thousands of devices working sequentially and/or in parallel on products. On the other hand, different manufacturing systems can have a large range of communication requirements with different amount of emphasis on throughput, delay, jitter, loss, and reliability. The communication requirements of manufacturing systems are outlined in Table 4.

Network architecture refers to a broad framework for organizing network communications and computation across end-points, relays, gateways, storage and compute resources available. Ultra-low latency wireless network architectures are likely to be differentiated from traditional network architectures due to unique application-specific requirements that emerge from different domains. At the same time progress in understanding these application-specific requirements and the architectural principles that can best support them can potentially yield a more unified approach that can be flexibly optimized and adapted to specific applications as needed. Section 3 addresses both the needs and guiding principles for architectures that support telesurgery, manufacturing, and virtual reality.

Section 4 addresses important technology aspects that are crucial moving forward for the design of the next-generation networks. The discussion in Section 4 considers both local area communication and long-range communications. For each case technology bottlenecks and challenges are identified, followed by a discussion of opportunities for future development. Finally, testbeds and demonstrators that will be essential to quantify achievable performance in realistic application scenarios are described.

There was general consensus at the workshop that improving latency in wireless networks is critical for enabling emerging mission critical applications that depend on consistent low latency. Moreover, opportunities exist for latency reduction at various levels of the protocol stack. Thus, now is an opportune time to invest in research toward the design of low-latency wireless networks.

1. Introduction

The workshop on Ultra-Low-Latency Wireless Networks was held on November 3-4, 2016 in Arlington, VA. The workshop's attendees included over 30 participants from Academia, Government and Industry, representing a range of perspectives: from network architecture and algorithms, to emerging applications. A particular focus of the workshop was on establishing communication requirements for the emerging applications of Virtual Reality, Telemedicine, and Automated Manufacturing. As such, the workshop involved a number of participants from both industry and academia with expertise in these application domains.

The first morning of the workshop was focused on establishing the communication requirements of these emerging applications. The application domain experts made presentation on the respective applications, and helped the workshop participants establish application requirements. These requirements are described in Section 2 of this report. In the afternoon, the workshop participants focused on identifying potential architectures for meeting the communication requirements. Led by experts in network architecture, the workshop participants developed strawman architectures for meeting the communication requirements of the three application domains under consideration. These architectures are described in Section 3 of this report. Finally, during the second day the workshop addressed technology limitations and challenges. Led by experts from industry and government, the participants examined the limitations of existing communication technologies, and identified opportunities for improving latency. These issues are described in Section 4 of this report.

The workshop was organized by Prof. Eytan Modiano (MIT), with support from Prof. Kyle Jamieson (Princeton), Prof. Ashu Sabharwal (Rice), and Prof. Sanjay Shakkottai (UT-Austin), under the sponsorship of the National Science Foundation (NSF). This report summarizes the findings of the workshop.

2. Application requirements

During the first morning of the workshop, participants discussed in detail three emerging application areas that would greatly benefit society, should ultra-reliable, low-latency wireless communications become a reality. These discussions focused on challenges and opportunities within each domain area.

The application areas considered were:

1. Telesurgery/telemedicine;
2. Augmented reality (AR) and virtual reality (VR);
3. Smart manufacturing.

We next elaborate on each of these applications, differentiating between classes of traffic within the same application, detailing their Quality of Service (QoS) needs, communication range, scalability challenges and power constraints.

2.1 Telesurgery/telemedicine

Telesurgery and telemedicine encompass several applications ranging from remote diagnosis to remote consulting during a procedure to true remote surgery. On the patient side, this application may require several cameras (some of which could be wearable by the local staff). On the remote expert/surgeon side, these applications may require a VR interface in order to provide an immersive sense of experience.

2.1.1 Use Cases

- Remote Diagnosis: In areas such as gerontology, dermatology, and physical therapy, an expert can observe a patient in a remote clinic/hospital (*e.g.*, in a rural area) and provide a diagnosis/treatment plan. On the patient side, this application may require several cameras (some of which could be wearable by the local staff). Since there is no actual feedback beyond observations and relatively simple requests by the expert, this application has the least stringent requirements in terms of delay.

- Remote Surgical Consultations and Support:
    - Planned - An expert can observe a planned surgery that takes place in a remote location (rural area or field hospital) and provide advice and guidance. This application has more stringent latency requirements, but not necessarily ultra-low latency ones.

    - Emergency - Complex life-saving procedures following an accident, injury, or a sudden health emergency cannot wait until the patient is transported to the hospital. These scenarios include cases where the ambulance and/or emergency medical technicians (EMTs) are far from the hospital, and emergency situations occurring in the roadside, sea, air, and military fields. In all these cases, a local medic could work with a surgeon via 2D or 3D telepresence to get support for

complex procedures. This application has similar requirements except most of the communication could be wireless.

Robustness is extremely important in both cases above, especially if the user (local caregiver or EMT) is dependent on the system. In such cases, end-to-end application level failure of over few seconds can affect treatment, and may have catastrophic effects.

- Telesurgery:
  - Local - An expert surgeon physically located near the operating room can participate in a surgical procedure, with the advantage that this surgeon does not need to scrub in and out, and thus can rapidly move between a few patients. In this case there is a need for a very tight feedback loop and so all the wireless communication is short range. The state of the art is the system provided by DaVinci Surgery (http://www.davincisurgery.com) which is currently mostly wired, and does not provide force feedback.

  - Remote - Remote surgery in rural areas may serve patients who cannot access hospitals. There need to be some care facility in the remote location for after care of the surgery. There is a need for remote follow up by the surgeon. The requirements are similar to the local scenario above, but should be supported over longer end-to-end distances. The end-to-end link could be wireless or wireline. It would be impossible to tolerate failures.

Because remote surgery has the most stringent requirements with potentially the highest impact, the following networking needs/limitations focus on that application.

2.1.2 Networking Needs/Limitations

Telesurgery has three main traffic classes: haptics, video (used for 3D geometric reconstruction), and audio. 3D video can be collected from 4K streams generated by 20 to 50 cameras; some of them (e.g., wearable) have to be wireless. Network capability permitting, some streams may get priority based on the focus of the surgeon and remote caregiver.

In all the above cases, the first hop from the wearable cameras, if any, may be a wireless link. Also, if a temporary setup is created in remote setup, the first hop may need to be wireless as well. Also, the connection to the remote surgeon's VR-headset requires wireless connectivity. The rest of the flow can be offloaded in wired connection for backhaul. However, in a disaster or accident scenario, all the hops might need to be wireless connection.

The requirements in terms of QoS metrics are provided below. In short, haptic has the tightest delay requirements (i.e., at most 10 ms), and video has the tightest throughput requirements (approximately 1 Gb/s).

	Haptics	Video	Audio
Latency	1-10 ms	20-50 ms	100 ms
Jitter	10 ms	30 ms	*ca.* 50 ms

Throughput	Negligible	1 Gbps	Negligible
Range	Up to 200 km		
Loss	1e-5 packet loss from the surgeon to the robot. Higher loss possible on the other path. Losses for video depend on compression.		
Power	Not a significant constraint except for wearable devices.		

Table 1: Networking needs/limitations for different telesurgery modalities.

Range: Because of the haptics requirement and speed of light constraints, 200 km appears to be the maximum range that can be supported.

Scalability: Scalability issues exist, especially in catastrophic events and multiple wearable cameras. Also, in case of multiple rooms, numerous cameras may stress the system.

Power: Power requirements do not appear to be a major issue as deployment scenarios allow that all the equipment can be either powered or recharged in a timely manner.

Failures: Rare failures in remote consultation may be tolerated, whereas remote surgery requires an extremely robust system. Although video reconstruction can tolerate some packet loss, 3D model reconstruction requires stringent packet loss rate. While in VR applications where this will result in noticeable errors, in telesurgery applications this may have catastrophic consequences.

2.2 Augmented and Virtual Reality

Augmented Reality (AR) adds supplemental elements (computer-generated) to the user's viewpoint, while Virtual Reality (VR) creates the entire scene the user sees based on multiple sensor sources (*e.g.*, cameras and other sensors). Both technologies operate in real time.

2.2.1 Use Cases

- Single-user AR/VR: AR head-worn systems such as Microsoft's HoloLens and smart helmets from DAQRI that provides situational context in the workplace help to improve both workplace efficiency and safety. VR/AR can also provide augmented work training, which is easier to follow and understand and reduce the training time and error. For the safety applications, latency is highly critical, while for VR/AR based training, latency may not be critical.

- Collaborative AR: Smart helmets from DAQRI are meant to provide more effective way to share information in workplace. For example, all personal perspectives can be sent to cloud to create a 360deg view and share any chosen perspective with each user in the same time.

- Virtual space/ group VR: One application of collaborative VR is VR meeting room so that people at different locations can hear, see and interact like they are in the same room during the meeting. In such applications, the latency is very critical due to the real-time nature of the interaction.

2.2.2 Networking Needs/Limitations

QoS: In terms of networking needs and limitations for AR/VR, we identified two basic scenarios and their requirements: (1) transmitting a single HD streaming with resolution *ca.* 1920x1080 and video frame rate 90 Hz in AR/VR applications and (2) sending uncompressed video and model to edge/cloud for processing. A key performance metric is end-to-end application level latency, which we define as the delay between a user action (movement of head) and the corresponding update of the display. This latency includes not only network latency, but also the effects of sample rates of sensors and processing delays. While it is possible to hide some of latency from the user by doing motion prediction, i.e., predict where the user will be looking when the frame is rendered, we generally would like to see user-perceived latencies to be 10 ms or less, with jitter of 1 msec or less, as shown in Table 2 below. Keeping both communication latency (milliseconds) and jitter (< milliseconds) low will make it easier to meet this goal and improve performance.

	Single HD Streaming	Uncompressed Video/Model
Latency	10 ms	1~10 ms
Jitter	<1 ms	<1 ms
Throughput	6 MBps	1 GBps
Power	For local computation (*e.g.*, rendering), 15-30 W is expected. When the computation can be moved to edge/cloud, the power can be reduced to 5 -10W.	

Table 2: Networking Needs and Limitations in AR/VR applications

Scalability: Density is likely to be highest in spaces such as rooms dedicated to telesurgery. Both sensors and need for wireless communication will be high density, and sensor density will depend on complexity of the scene (*e.g.*, number of people and objects) in order to obtain views of all aspects of the scene. In terms of hardware, for single-user VR, a single room may need one to 100 cameras depending on the application. For collaborative applications and virtual space, the requirement can be expressed in terms of devices per person. For example, an application may require four cameras per person. Availability, connectivity and coverage are also important factors that decide the scalability of VR/AR.

Another dimension of scalability is whether the technology can scale to a large user population. A critical factor that decides this aspect of scalability of VR/AR is cost. We expect that for the consumer market, a price point similar to a smartphone may be acceptable for a VR/AR device.

Level of importance of different data types: Within a single VR/AR application, different types of data will have different levels of importance, which will determine the resource allocation of the spectrum. For example, audio data which is at a low data rate in general has higher importance than video. Special sensors such as sensors for tracking eye movement, for object quality of experience measurements, and thermal sensors for firefighting applications have the highest importance in the corresponding applications. Other sensors such as ultrasound/IR may have the lowest importance. The levels of importance are summarized in Table 3. We also discussed an application-independent method to quantify the importance. Since data are generated with different sampling rates, a low sampling rate is an indication of low importance. Therefore, one method for identifying the levels of importance is to decide based on the sample frequencies.

Data type	Special Sensors	Audio	Video	Other sensors
Importance	high	→		low

Table 3: The Level of Importance of Different Types of Data in AR/VR applications

Computation: A trend in AR, as devices become more personal, is decreasing user tolerances for carrying a given size form factor on their person. One promising direction is to push computation to the edge of the network, hence the environment has the compute capability, and can for example stream information to users. A key question here is where exactly the computation itself should reside.

Failures: Depending on the application, the impact of a failure varies from unnoticeable to disastrous.

- Voice/Video: transmitting audio and video, users may not even notice a few packet losses.
- 3-D Model Reconstruction: loss of few packets can significantly impact the correctness of the 3-D reconstruction, which becomes noticeable to users.
- Special sensors/mission critical: For critical applications such as some industrial applications, it can lead to a disaster situation. Therefore, for critical applications, failure should be extremely rare, which requires reliable and ultra-low latency wireless networks.

2.3 Manufacturing

Manufacturing systems arise in such diverse domains as automotive/aerospace manufacturing, high-speed semiconductor manufacturing, smart-grid control, additive manufacturing, etc. Accordingly, the networking needs of these systems can range vastly in their scale and requirements. On the one hand, the scale can range from a single or a small group of co-located devices to a large plant that encapsulates thousands of devices working sequentially and/or in parallel on products. On the other hand, different manufacturing systems can have a large range of communication requirements with different amount of emphasis on throughput, delay, jitter, loss, and reliability.

2.3.1 Use Cases

To paint a clearer picture of various manufacturing system requirements, we can organize these systems into the following three service categories based their functionality:

- Control: This functionality is aimed at high-fidelity closed-loop control of devices.
- Diagnostics: This functionality is aimed at gathering and monitoring of ongoing manufacturing activities
- Safety: This functionality is aimed at preventing violation of preset constraints or other anomalies that indicate safety-critical events.

2.3.2 Networking Needs/Limitations

QoS: The above categorization of functions allows us to present in Table 4 the relative differences between their service requirements in terms of its throughput, delay, regularity of service, sensitivity to losses, and reliability.

Functionality	Throughput Requirement	Delay Sensitivity	Regularity Requirement	Data Loss Sensitivity	Reliability Requirement
Control	Medium	High	High	Medium-High	High
Diagnostics	High	Medium	Medium-High	Low-Medium	Medium-High
Safety	Low	Very High	Low	Very High	High

Table 4: Networking needs and limitations of the Control, Diagnostics, and Safety functionalities of manufacturing systems with respect to various key communication performance metrics.

While Table 4 is helpful in providing a relative comparison of service requirements of different functionalities in most typical manufacturing systems, it is also worth discussing these items in further detail under specific instances to present some absolute values and highlight some exceptions in some cases.

- *Safety-oriented* communication demands occur in the rare instance of imminent danger to humans and/or to the functionality of important devices. These events may be triggered by either automated safeguards, such as breaching so-called “light-gates” that guarantee minimum proximity of humans to danger zones, or human-signaling to prevent an imminent danger detection. As such they happen rarely and unpredictably with low data load (as low as a few bits), but when they do happen must be communicated reliably almost instantly, preferably within a few hundreds of microseconds, and not more than 1 millisecond.

- *Control-oriented* communication demands can happen at various scales of time and space based on the level at which they are implemented. We can broadly divide these scales into: *hyper-local* that is primarily concerned with the internal control of a single device; *local* that is concerned with the control and coordination of a group of co-located devices; and *remote* that is concerned with the control of devices from a distance. In principle, the delay sensitivities, reliability requirements, loss sensitivities, and regularity requirements get sharper as the controller goes from remote to hyper-local. For example, delay

requirements can be in the order of one millisecond or even less for hyper-local precision motor control in semiconductor manufacturing. In contrast, delay requirements for local coordination of co-located robotic control systems can be tolerable in the range of 10-100 milliseconds. Yet, looser delay requirements of few hundreds of milliseconds can be acceptable for remote control of devices such as thermostats in smart grids.

- *Diagnostic-oriented* communication demands differ from the previous two types in that it typically has a *human-in-the-loop*, and therefore is constrained by the biological limits of human operation, such as less than 100 milliseconds of delay being indiscernible for most purposes. Moreover, diagnostic services are primarily for monitoring and analyzing manufacturing system operation, and are typically insensitive to delays even above one second and to data losses unless they are significantly large. On the other hand, the amount of data load can be high due to the aggregation of data from many devices.

Range: The range of communication can range from centimeters in the case of hyper-local semiconductor manufacturing, all the way to hundreds of meters in the case of remote control and diagnostics. Most typically, communication range is expected to be in the order of meters.

Scalability: The scalability requirements will vary in different manufacturing systems. It will be particularly important in local control of many devices, such as in the coordination of co-located robotic devices or as in the control of nozzles in semiconductor manufacturing.

Power: In several manufacturing systems, such as controlling motors, the ability to leverage easily-deployable and wirelessly-connected sensors can increase efficiency and reduce maintenance costs. In such cases, energy-efficient use of limited battery capacity will be critical for longevity and performance. Such conditions emerge also in other manufacturing systems where the devices are mobile or when the presence of cables affect the operation of the device.

Failures: Manufacturing systems can be highly sensitive to communication failures, both for obvious safety concerns, and also for production efficiency. In particular, *time-synchronization* is typically critical in many manufacturing applications across all three functionalities. In the absence of reliable time-synchronization, many operations may fail due to the importance of temporal dynamics of manufacturing processes. To avoid such undesired phenomena, current practice is to perform over-sampling of system states at 5-10 times the required Nyquist rate. However, this inefficient practice puts a heavy burden on the traffic load, and must be reconsidered during the transition from wired to wireless backbone. There is a need to guarantee low delays for such traffic at smaller throughputs than those used in practice today, which cannot be resolved through the use of the currently prominent wireless standard IEEE 802.11n.

3. Architecture

Network architecture refers to a broad framework for organizing network communications and computation across end-points, relays, gateways, storage and compute resources available. Ultra-low latency wireless network architectures are likely to be differentiated from traditional network architectures due to unique application-specific requirements that emerge from different domains. At the same time progress in understanding these application-specific requirements and the architectural principles that can best support them can potentially yield a more unified approach that can be flexibly optimized and adapted to specific applications as needed.

In the sections that follow, we discuss both the needs and guiding principles for architectures that support three specific applications: telesurgery, manufacturing, and virtual/augmented reality.

As discussed in previous sections, each of these applications (telesurgery, virtual reality, and automated manufacturing) have specific network requirements, including throughput, latency, jitter, security, reliability, light-weight control, computation, and power usage. These requirements in turn translate to specific architectural requirements. For example, in telemedicine, the architecture must address both a backhaul component of the network that connects the doctor at a hospital to an ambulance in the field and the local network in the ambulance providing connectivity for wearable cameras and sensors on the first responder and patient; in manufacturing, there is a need for the architecture to be designed so as to support a very heterogeneous mix of traffic requirements with relatively low-rate data streams for real-time automated control being combined with high bandwidth low-latency streams for remote users to monitor and operate equipment in real time; while in virtual/augmented reality applications, the architecture must provide support for computation needed for feature extraction, model fitting, and rendering and in some cases will need to provide greater support for multicast flows connecting individuals located far from each other.

These architectural needs lead to application-specific principles spanning many dimensions, including how specific needs translate into timescales and spatial scales of interactions, what elements should be secured and when, what types of interactions are needed and what can be autonomous, distributed versus centralized and hierarchical approaches, address centric and information centric architectures, communication modalities (e.g. data rates and link-level properties, multi vs single band, multi-hop vs point to point), and how other capabilities such as storage and computation are integrated into the network.

While each of these architectures have unique elements, there are certain fundamental needs and principles that are common to all these applications, all stemming from stringent requirements of ultra-low latency. For example, interactive VR (with applications both in industrial workplaces and telesurgery) as well as emerging additive manufacturing technologies need both low latency and high rate. More generally, all three domains require a highly agile network architecture that can adapt and respond to the heterogeneous QoS requirements of different applications. This may be realized, e.g., through virtualization or resource slicing, where the slicing itself needs to be highly flexible and adaptive. Safety-critical traffic may be allocated dedicated resources. And for all the applications, the network architecture will need to adaptively serve for a range of QoS requirements measured through a mix of latency, low-jitter, throughput, and loss

requirements. This means that we need to reconsider the networking architecture to fit the latency-sensitive and time-varying nature of the wireless conditions and the QoS requirements of the traffic at the time. Finally, all these three ultra-low latency wireless networking applications typically involve a human-in-the-loop, which brings unique challenges but also potentially opportunities to incorporate interactive human inputs and feedback about performance into the architecture.

3.1 Telesurgery

3.1.1 Architecture Needs

As discussed previously, there are a range of different telesurgery use cases. In this section we focus on one of the more demanding applications from a networking perspective: performing remote surgery with the aid of a first responder in an ambulance. This scenario may involve mobile units, for instance, AR headsets worn by surgeons and first responders and remotely-controlled telesurgical equipment deployed with first-responders (e.g., those in ambulance). Providing connectivity to such mobile units clearly requires a wireless solution. Further to provide the real-time communication needed for telesurgery, low latency communication is required. If immersive AR is used to enhance this, an architecture that supports high bandwidth/low latency video as well as low latency haptic and audio communication is needed.

The architecture of this application naturally separates into a backhaul component that provides connectivity from a hospital to the ambulance and local access within the ambulance for providing connectivity, e.g. to devices worn by the first responder. We emphasize that the overall end-to-end performance must be managed to meet the application needs discussed earlier.

It would be desirable for such a system to operate in a number of different geographies including both urban and rural, and so the architecture must be flexible enough to accommodate different propagation environments and different available infrastructures.

Lack of security and reliability in telesurgery can in the worst case have catastrophic consequences, or run into legal concerns if a procedure goes wrong – security measures to ensure up-front authentication and preventing denial of service are essential.

3.1.2 Architecture Principles

To provide the needed high-bandwidth, low latency backhaul an architecture is needed that can leverage a range of heterogeneous communication technologies including mm-wave and traditional cellular connectivity. Part of this backhaul may include reaching an access point and then using wire-line backhaul. In this case, the wire-line network would need to ensure the needed latency requirement is met while the network is shared with other applications, for example by using network slicing and appropriate resource reservation algorithms. In some cases, simultaneously connecting over multiple technologies may be required.

The local connectivity might be provided using a WiFi-like technology within the ambulance, but again, techniques would be needed to ensure that the needed quality of service is met (e.g. to avoid excessive interference). The ambulance also provides a platform for local power and computation that can be exploited to reduce the demands on wearable devices and perhaps reduce the needed communication bandwidth.

Given that the communication connectivity can vary with locations, the architecture must also enable system-wide optimization of sensing (e.g., 3D vision), networking, and computation; examples include, but are not limited to, viewpoint-adaptive optimization of 3D data streaming (such as scalable video coding and localized scene streaming) which could be used to reduce the needed transmission rates. Techniques for supporting prioritization of traffic (e.g. control commands over video) would also useful. Approaches for ensuring security and reliability will need to be an integral part of the architecture. Again, leveraging multiple technologies can aid in this.

Somewhat distinct from other applications, power may not be a major concern for telesurgery as the ambulance can provide a ready power source (with the exception of wearables). Unlike the safety- and control-related communication in manufacturing plants, the communication distance may be long (e.g., up to 200km) in telesurgery, and the bandwidth requirement may be higher (e.g., requiring 125Mbps bandwidth for a telesurgery system using 25 UHD cameras). The nature of human-in-the-loop (instead of simply machine-to-machine) and the different impact of packet loss on different signals (e.g., haptic, visual, audio, and control signals) are also key characteristics of telesurgery.

3.2 Automated Manufacturing

3.2.1 Architecture Needs

Low-latency wireless connectivity is a critical emerging need in manufacturing environments. Currently, majority of the connectivity, e.g., between sensors, actuators and controllers, in manufacturing environments are realized using wired networks. One example is the manufacturing of sophisticated machines like a jet engine, which can involve hundreds of wires and cables connecting related equipment. Replacing the wires with wireless links can bring substantial benefits, in terms of: (i) *Reducing the wire hazards*: wires cause various hazards for human operators and robots, e.g., tripling, short-circuiting, electricity leakage; (ii) *Reducing manufacturing cost*: diagnosing and maintaining (replacing and repairing) the wires can incur non-trivial cost; (iii) *Reducing failure rate*: wired connection can fail as they get worn out; and (iv) *Flexibility*: it is much easier for wireless links to be rerouted and reconnected.

Low-latency is needed in both the low-rate control and high-rate data transfer scenarios. For example, coordination of the sensors/actuators requires tight synchronization, and hence latency-guaranteed wireless communication. On the other hand, there exist high bit rate interactive applications. One example is in scenarios that require tight synchronization between multiple devices that are coordinating together in specific tasks, e.g., multiple mobile robotic platforms that are coordinating together in moving a large palette from one location within the plant to another. Another such example, is in synchronizing real-time 3D scenes to a cloud backend for augmented reality inside industrial environment.

Of course, in addition to such low-latency communication between devices and other system components, there are other scenarios where a high throughput path would also be desirable, e.g., video from a manufacturing plant floor is streamed over to cloud-hosted storage for analytics or diagnostics.

Based on these observations, we believe that in a manufacturing environment, the network architecture must be able to support two types of services:

(i) Low latency and high reliability for delivery for a small fraction of traffic e.g., for real-time control and coordination of multiple devices, or for implementing some safety functions. The aggregate data volume of such traffic may be low, but when desired, such traffic needs assured delivery within some small time windows.

(ii) High throughput applications from a relatively small number of devices. An example of such traffic is a drone which may need to deliver real-time video to a central site. A factory plant may have tens to hundreds of cameras sending real-time video data to a central control center. A large number of 3D printers may be co-located in a manufacturing site, and each would need a camera to monitor its progress in real-time.

To satisfy these requirements, the network architecture must be scalable with respect to the number of nodes connected. This is critical because the availability of reliable wireless connectivity will in turn trigger more and more connected devices. For environment with deterministic traffic (e.g., periodic sensor data upload), resource can be reserved a priori, so the network can easily scale. Further, the traffic within a large manufacturing plant may tend to have a periodic structure for much of the communication, but with limited amounts of unpredictable and asynchronous traffic exchange.

The network itself needs to be highly agile. It should have the ability to prioritize safety-critical data. It should also have minimum guarantees on delay, throughput, regularity of updates that are absolutely required for control and diagnostics. It should adapt its configurations between these extremes as the latency-sensitive QoS requirements and wireless resources fluctuate. Finally, security is another factor that the network architecture needs to handle. Wireless network is broadcast in nature, and so, it is vulnerable to eavesdropping and jamming attacks. Further malicious attackers maybe motivated to hack into these networks and modify the behavior of machines and devices. Given the cost and criticality of manufacturing systems it is important for the low-latency network architecture for manufacturing incorporate necessary safety measures from the get-go.

3.2.2 Architectural Principles

To meet the goals described above, it is envisioned that the architecture will support heterogeneity in communication. There will be at least two categories of traffic --- some that require ultra-low latency but are usually low bit rate; and others that have elasticity in latency but depend on higher bandwidth. Providing some guarantees for the first category of traffic is critical to ensure safety and real-time coordination that are essential to the natural functions of this environment.

Further, the environment itself is naturally hierarchical. Communication needs can be broken into three types --- hyper-local, local, and remote. Further, given that the entire environment is usually under a single administrative control, there is an opportunity to create a more scalable and efficient communication structure. In particular, from the network topology perspective, hierarchical network architectures may be needed to support connectivity at the wide-area level, the cell level, and the machine level. At the local levels, the architecture must enable ad-hoc communication between various local devices for faster and real-time interactions.

Given the critical nature of security in this environment, appropriate mechanisms are required to protect external hacks and wireless-centric attacks, including various forms of jamming. Certain mechanisms, such as frequency-agile communication, may be built in the networks to ensure certain level of protection. Moreover, the architecture must incorporate security from the ground up, for example requiring strong authentication of nodes on the network to prevent malicious individuals or organizations from stealing intellectual property from within the manufacturing system or injecting malicious code to disrupt critical operations.

Finally, for manufacturing applications, the network has to adopt a hierarchical architecture, with a mix of infrastructure and D2D connections. In addition, a narrow spectrum of low-latency, low-rate, but safety-critical applications unique exist in the manufacturing environment. From the hardware and communication level, manufacturing environment suffers more from narrowband noise, interference, and severe obstruction by metal objects. The wireless devices themselves may have very tight form-factor constraint, in order to fit together with the small sensors or other manufacturing components.

3.3 Augmented Reality and Virtual Reality

3.3.1 Architecture Needs

The key aspect of AR/VR applications is the need for rich, timely, and accurately positioned scene rendering for a user, or a group of users. Failure in this respect leads to lower user Quality of Experience (QoE) and cyber sickness. Wearable displays are limited by their processing capabilities, weight, battery power, and to some extent heat dissipation allowance, and so may need to offload the rendering functions to nearby, more powerful hardware. The ability to offload rendering, however, faces network challenges specific to the scale and scope of an AR/VR application.

Additional aspect of AR applications is the need for real-time streaming of videos and other sensor data to bring the personal view: 1) into a broader perspective automatically, and 2) to communicate with remote experts for interaction. In the first case, one can use video streaming to register local perspective to a global map based on GPS and image matching. This is very helpful when multiple users of AR wearable need to collaborate. Such data streaming can also be used to refine and update existing global map, critical for work planning and execution.

In hyperlocal scenarios, within a single room, or a vehicle, rendering offloading will rely on *single hop* connections between wearable displays and rendering hardware. Latency should be under 100 microseconds, and so communication may take place, for example, over Layer 2 connections to a rendering process attached/incorporated into a WiFi access point, or a cellular base station.

In local scenarios (within a factory floor or a shopping center), rendering offloading will rely on two or more hop paths, where the wearable display and rendering hardware are connected to the same access point/base station over Layer 2, or Layer 3. Latency should be under 1 millisecond at a distance of 10-100 meters. Rendering functions in local scenarios may take advantage of Mobile Edge Computing (MEC) nodes attached to network hardware.

Finally, in long range scenarios, at distances greater than 100 meters, rendering offloading will rely paths over multiple Layer 3 connections, possibly over multiple lower layer technologies (wired and wireless) and administrative domains. Latency requirements are under 10 microseconds. Long range scenarios may also extend to connect groups of users located in different cities through telepresence. While in this scenario challenges of providing low latency at the different layers compound, successful technologies, such as OnLive and Outatime, are in principle able to take advantage of aggregation of rendering hardware for lower system costs. Proposals for ultra-low latency intercity networks (multihop mmWave, LEO constellations) may provide sufficiently low latency end-to-end paths to meet latency constraints.

Architectures for AR/VR should also consider group communication patterns. For example users may share context in a same physical location and also share network resources. In such scenarios multicast communication solutions should be considered to reduce communication bandwidth requirements and threats to determinism of communication latency stemming from congestion/interference.

AR/VR applications can communicate real-world user actions, which may create threats to user privacy. Architectures that support these services must consider these threats in the context of local laws and social norms.

3.3.2 Architectural Principles

Several architectural decisions are key in order to achieve the aforementioned needs:

Location of computing nodes: One important principle is where will the computation needed for augmenting reality take place. Lightweight headsets with limited power supply may only perform little or no rendering. Full rendering can be offloaded to a node in the same room if the wireless latency is below 1ms and a 3D model computation can be executed even further away. Possible location of computing node will depend on parameters (bandwidth, latency, jitter) of particular communication links and on specific application requirements (we may require local rendering only for particularly high-quality experience).

Spectrum and interference management: This is a general issue in any wireless system, but particularly pronounced in the case of AR/VR scenarios because of its extreme demands on bandwidth and latency. Different radio frequencies offer different communication characteristics. Lower frequencies propagate further and provide range but in turn have less radio bandwidth available, hence link capacities are lower. On the other hand, more spectrum is available in high frequencies but only at a limited range. Similarly, different radio technologies are deployed in different spectrum and offer different means of interference management. For example, Wi-Fi networks are denser and can offer high bandwidth but interference from nearby transmissions can cause high jitter whereas cellular access could provide better jitter and latency but lower capacity. The systems will need to be aware of these characteristics, and manage traffic across available interfaces. Equally, we need to come up with new access and interference management schemes for wireless that will be specifically designed with low latency and jitter in mind.

Network management and quality of service: in long-range scenarios in which wired access networks are required, they are equally important to consider when accounting for the overall latency budget. The latency between two endpoints cannot be lower than what dictated by the speed of light, however, today's wired network can add substantial delay beyond the physical

lower bounds because of the way public Internet architecture is organized. We need to rethink the way wired network bandwidth is managed. One example is research on future internet architecture (NSF FIA and FIA-NP programs), which led to several proposals for inter-domain, flow based forwarding that lend themselves better to QoS support than forwarding based on BGP paths (e.g., Scion, Icing). Another example is recent work tries to leverage software defined IXPs to stitch together paths with QoS properties.

Application-level optimization: Even with improved network architectures, application requirements will occasionally exceed the available resources. Applications will need to be aware of that and manage it appropriately. One example is adaptive quality of experience. When bandwidth becomes limited, the application may decide to limit the video quality, avoid rendering textures, or render only a narrow field of view. Similarly, when latency increases, the application may decide to move the rendering to a head-set, again trading off with the user experience. Furthermore, accurate time-stamping can improve latency sensitivity by exposing timing information to the application layer. Finally, there is a need for an appropriate admission control when the required aggregate demand cannot be supported without significant quality of experience degradation.

Shared communication: For certain kinds of VR/AR applications, such as online classes or events, in which multiple users share the same view, part of communication and computation can be shared. One example is a network-level multicast. Another example is shared video rendering for multiple users that share an immersive experience in the same room. Architectures will have to be aware and exploit these opportunities in order to meet the strict requirements.

Security and privacy: Security and privacy are important parts of any information system design. These gets pronounced in case of lightweight headsets that are expected to be deployed as glasses and other accessories. Another important security issue is wireless jamming. The architecture should ensure that a safety critical system, such as AR in some industrial applications, is robust to external jamming by malicious users. AR/VR applications (with shared hardware) may store private information between sessions and the architecture design should ensure that data is stored in a way that does not violate privacy and that devices in the system are not accessed or tampered with by unauthenticated, unauthorized users.

Finally, many of these principles are common to other wireless low-latency scenarios, but some are unique to AR/VR. Security and privacy concerns are more pronounced in AR/VR scenarios because they often run on publicly accessible systems, unlike manufacturing and tele-surgery. Similarly, AR/VR applications will often need to support different deployment scenarios, such as dense areas (shopping malls, stadiums), and long range links (car passengers with headsets). These scenarios are prone to more variable wireless link qualities, and thus several architectural principles (such as spectrum and interference management and application-level optimization) become high-priority.

4. Technology Gaps and Opportunities

In this section, we discuss three important aspects that are crucial moving forward for the design of the next-generation networks. An important outcome of discussion in the previous sections is that there are distinct requirements for short-range communications (less than 10s of meters) and long-range communications (in 100s of meters or longer). Thus, the discussion in this section is organized into two subsections accordingly; Section 6.A discusses hyper-local and local area communication and Section 6.A discusses long-range communications. For each case, we first discuss technology bottlenecks and challenges, followed by opportunities for future development and finally, testbeds and demonstrators that will be essential to quantify achievable performance in realistic application scenarios.

We note that the need for low latency communication is also driving wireline technologies. The demand from Internet of Things, automotive networking and video applications are driving changes to Ethernet technology that will make it more time-sensitive. Key to those changes are a number of developing standards. An example is recent efforts from the University of New Hampshire InterOperability Laboratory that has set up three new industry specific Ethernet Time-Sensitive Networking consortiums -- Automotive Networking, Industrial Networking, and ProAV Networking aimed at developing deterministic performance within standard Ethernet for real-time, mission critical applications. Quoting Bob Noseworthy, Chief Engineer of UNH-IOL, "Standards-based precise time, guaranteed bandwidth, and guaranteed worst-case latency in a converged Ethernet network is a game-changer to many industries." Additionally, the Avnu Alliance is developing an ecosystem of low-latency, time-synchronized, highly reliable synchronized networked devices using open standards through certification.

4.1 Hyper-local and Local Area Communication

4.1.1 Technology Bottlenecks and Challenges

There are significant gaps between the requirements of ultra-low latency applications and the performance of state-of-the-art technologies in several fronts, including application-specific features, PHY and MAC limitations, and network system architectures.

Emerging applications that require ultra-low latencies can introduce new challenges beyond just latency requirements. Consider the application of manufacturing, which consists of thousands of sensors deployed within a factory. In such environments, even guaranteeing connectivity can be difficult. Moreover, many sensors are deployed in harsh environments that are highly reflective and absorptive in signal propagation, such as within a metal pipe or inside an injection molding machine. Finally, as nodes are not necessarily connected to power supplies, they need to be ultra-low-power, and may need to harvest energy from environments. This makes low-power communication a necessity in many cases. Also, local control and safety services come with high reliability and regular service requirements, in addition to ultra-low delays.

AR/VR also imposes additional unique challenges. To enable ultra-low-latency in end-to-end communication for AR/VR, it might be necessary to host some services or prefetch content at access points. This creates additional problems with such as buffer-bloating in queue management and TCP-compatibility. For ensuring stringent Quality-of-Experience (QoE)

requirements for AR/VR users, providing low jitter, regular service, and high-reliability in addition to ultra-low-latency, are also important challenges.

On the limitations of the PHY and MAC layers, it seems obvious that high bandwidth can enable both high throughput and low delay. However, different frequencies have significantly different behavior. For example, mmWave is a very promising band to enable ultra-wide bandwidth communications, but it suffers from serious attenuation and usually requires line-of-sight. To provide consistent ultra-low-latency performance to users, new mechanisms that aggregate information on multiple different frequency bands are necessary. As different radio access technologies (RATs) are used for different frequency bands, mechanisms need to be developed to integrate different RATs seamlessly.

Finally, there are several important shortcomings in existing network architectures. Current network designs only focus on optimizing the network performance without sufficient consideration of computation and best-effort delay performance. The overhead of computation cannot be ignored to deliver ultra-low-latency service. New ideas for the co-design of networks and computation are necessary. Also, in many applications of multi-user AR/VR, multicast is used to deliver common information to all users. Mobility and heterogeneous channel qualities of users can present further challenges, especially when using highly directional bands such as mmWave. On the other hand, in manufacturing, deployment and coverage of a large number of nodes can become a major bottleneck. To support the massive number of sensors, self-organizing network techniques that also ensures ultra-low latency become critical.

4.1.2 Opportunities

Recent advances in network optimization and adaptive control, network architecture, and physical layer techniques, the nature of short-range communication, and the availability of heterogeneous communication mechanisms and spectrum offer opportunities for tackling the aforementioned technical challenges.

There have been significant progress in mathematical network optimization and control in the past decade, which has provided the mathematical foundation for designing and reasoning about networks and which have provided distributed solutions for network-wide optimization. In recent years, progress has also been made in incorporating low-latency and other short-term requirements into the network optimization and adaptive control framework. For instance, there have been work that explicitly consider data delivery timeliness in the constraints and/or objectives of the mathematical models, and there have been tractable algorithms that achieve optimal real-time capacity while ensuring data delivery delay and regularity. Building upon these results, the research community is poised to answer fundamental questions, for instance, how to formulate the mathematical problems such that they reflect different application semantics (e.g., reducing deadline or deadline miss, reduce delay and/or jitter, probabilistic or deterministic delay guarantees), and how to optimize multi-scale performance metrics such as per-flow, long-term timely throughput and per-packet delay guarantee. In parallel with progress in network optimization and control, significant progress has also been made in network system architecture, and emerging network architectures such as information-centric networking can be leveraged to enable system-wide optimization of networking, computing, and application logics.

The locality of the systems of short-range communications also offer opportunities for lightweight yet tight coordination among nodes for network optimization and adaptive edge computing. In this

context, we can develop latency-aware queue management schemes to avoid issues such as buffer bloating; information-centric networking mechanisms that leverage heterogeneous communication networks and multiple, heterogeneous spectrum for ultra-reliable, real-time, high-throughput communication. We can also leverage emerging physical layer techniques such as mmWave, massive MIMO, VLC, and full-duplex communication to enhance network real-time capacity.

To address fast, micro-mobility in AR/VR and telesurgery (e.g., headsets) as well as manufacturing (e.g., robotic arms), context-aware mobility prediction can be leveraged to optimize network control such as beamforming, interference control, and scheduling. To enable scalable real-time communication in dense network settings, multicast and multi-scale resource management (e.g., allocating resource over time and space) can be leveraged. For low-power communication (e.g., in manufacturing), energy harvesting techniques (such as backscatter communication and motion-energy harvesting), low-power, low-rate waveforms, and low-power real-time network techniques can be developed.

4.1.3 Testbeds and Demonstrators

To demonstrate the efficacy of new design and architectures, the testbeds should support multiple new features. First, the testbed should allow precise measurement of latency and throughput, which is crucial to evaluate the performance of new designs for low-latency wireless network designs. Considering that the lowest latencies demanded by the applications can be as low as sub-millisecond, synchronization protocols like PTP and WhiteRabbit can be utilized for synchronization of multiple nodes in a testbed with many nodes.

Second, for short range communications, many bands can be utilized. Thus, it is desirable for a testbed to support multiple forms of communication bands simultaneously. The support should allow flexible use of the bands at all network layers, including the physical layer.

Third, for applications such as manufacturing and AR/VR, there is micro or local mobility; for example, movement of heads or arms. The testbeds should support repeatable mobility to enable experimentation for different application use cases. The repeatability could be achieved with robotic platforms, where captured human motion traces can be used to emulate application use cases.

Fourth, for each application domain, availability of benchmarks could facilitate comparisons of different wireless designs and their impact on the performance of each application. For example, different AR/VR use cases could be captured as example scenarios along with metrics to define application-specific performance. Another advantage of benchmarks will be to avoid human-in-loop experiments and, in the process, allow automation of performance testing.

Finally, modularity of testbeds will be highly desirable, to allow using common components for different application scenarios. Modularity will also ensure that optimizations from one application scenario can be used for other application demonstrators.

The demonstrators could be at three stages of development. The first stage will include targeted demonstration of specific concepts, for example, transmission of information over diverse spectrum that could “aggregate” sub-6 and above-6 GHz spectrum. The second stage will include system level demonstrations using emulation benchmarks, thereby building confidence for

system level performance and could include end-to-end performance evaluation of latency and its jitter. The third stage could include testing in actual scenarios but with high levels of control and safety precautions. For example, an emulated factory floor could be used to test multi-sensor networked control applications.

4.2 Long-range Communications

4.2.1 Technology Gaps

As outlined in previous sections, applications being discussed (VR, telemedicine and manufacturing) present a rather challenging set of bandwidth, latency and reliability demands. These demands are not met by the current technology for long-range communication, as discussed next. Available and discussed technological solutions extend beyond cellular networks (LTE 4G and 5G) and include unlicensed spectrum solutions such as LoRa, Weightless, SigFox and other low-power wide-area networks (LP-WANs). The range of frequencies presents both a challenge and an opportunity: LP-WANs are typically used at sub-6Ghz range, while 5G is envisioning active usage of mmWave (30Ghz and above). These ranges present rather natural usage patterns: sub-6Ghz could be used for low-rate, long-range (single-hop) data, such as diagnostic and control messaging for manufacturing applications, while going to mmWave is essentially inevitable for satisfying the high-rate demands of video streaming. The challenge lies in designing architectures permitting multiple-interface communication. A single protocol stack, ideally, should be able to distribute the load between available physical interfaces, but such solutions are not currently available.

Another crucial bottleneck comes from the energy-per-bit considerations. In LP-WANs it is customary to restrict data rates to a few kilobits per second. This restriction is dictated by a simple calculation: given the total bound on radiated power (in Watts) and propagation loss over a 10km range, one needs each bit to occupy significant time on the channel for the receiver to be able to collect enough energy for meeting fundamental (information-theoretic) requirements for Eb/No. Clearly, such low communication rates are incompatible with low-latency requirements. Possible solutions will involve multi-hop networking, directional antennas and working with the licensed spectrum to increase the limits on radiated power.

A significant issue with the current technology is going to be scaling to the orders of magnitude larger number of users. In manufacturing it is customary to have 100,000 or more sensors in a single plant. Servicing this many devices is outside of reach of current solutions in either the licensed (LTE) spectrum, or unlicensed spectrum (WiFi and LP-WANs). As an example, wireless networking on stadiums is provided by installing tens of thousands (!) of access points. While this (to some extent) solves the problem, a long-range solution would not only be more attractive economically, but also extend to other scenarios (such as servicing large crowds on city streets in spontaneous or emergency situations).

The final issue with current technology is the scheduling and QoS. It is strongly desirable for the users to be able to subscribe to access plans with carriers that would guarantee a certain minimal QoS in terms of (bandwidth, latency, reliability) triplet. Such subscriptions are not currently available for various reasons, among which the absence of good pricing mechanisms was mentioned, and the absence of appealing cross-layer QoS information. It is also important to notice that in the end, what matters to the end user is the end-to-end perceived distortion (in terms

of image quality and perhaps latency), also referred to as quality of Experience (QoE), so it is important for the applications to be able to provide mechanism for informing the bottom layers of network stack on how the target distortion metric deteriorates in terms of each of the (bandwidth, latency, reliability) triplet.

4.2.2 Opportunities

In bridging the various technology gaps to enable low latencies services like tele-surgery, automated manufacturing, and mobile AR/VR over long distances, various opportunities for research have been identified.

Spectrum Management: The need for increased bandwidths is motivating the use of higher frequencies like mmWave, which faces challenges in terms of increased attenuation and link sensitivity for providing reliable, mobile access. It is also critical to innovate in the unlicensed spectrum (eg, 3.5 GHz, 5 Ghz), which can lead to innovative mobile services from green-field operators without having to rely on mobile carriers. Given the availability of diverse spectrum and their appropriateness for different service requirements, intelligent spectrum management solutions that make the best, aggregated use of available spectrum options are needed. This requires the design of multi-homing solutions that can manage highly disparate interfaces (frequencies) and their access technologies at fine time-scales to deliver the desired low latencies.

Access Techniques: Wireless access in today's mobile networks are rather rigid and incapable of catering to low latency requirements. Wireless access needs to adopt a more flexible transmission structure, whereby the transmission slots can be made adaptive and finer (order of microseconds to milliseconds). Multiple access techniques that are capable of scaling to technologies like massive MIMO are needed, while keeping the control overhead minimal so as to deliver the low latencies. The design of error correction codes for low data rate (hundreds of bits to few bytes), but time-critical IoT traffic needs to be re-visited. Finally, very little application information is leveraged in wireless resource management today. Even simple information relating to end-device capabilities (e.g., location, interface capabilities, etc.) are not leveraged by wide-area wireless networks to better optimize their use of spectral resources. Going forward, tighter cross-layer interactions between wireless access and applications is needed, so that they can be jointly optimized to better serve the application requirements.

Deployment models: Network deployments for serving low latency applications raises several interesting challenges. Bringing base stations or access points closer to the end devices through the deployment of small cells is promising from the perspective of providing increased data rates and low latency. However, this faces the challenge of finding a cost-effective backhaul to connect these small cells. In some scenarios, it may be viable to deploy a multi-hop architecture, with a combination of fixed-access mmWave backhaul and lower frequency (sub-6GHz) small cells for access to accommodate user mobility. To provide ultra low latencies for certain applications like tele-surgery over longer distances may need dedicated wireless backbone infrastructures (e.g., using microwave backhauls, LEO satellites) that bypass the public internet to serve dedicated traffic.

Core Network Orchestration: While improving latencies on the wireless access through the above-mentioned approaches is critical, it is equally important to keep latencies in the mobile core networks low, which is not possible today. There are several opportunities to make this happen through the introduction of the concepts such as network slicing, virtualization and mobile edge computing. Virtualization and network slicing allow both the core and radio access network to

differentiate and handle traffic differently by providing dedicated resources for traffic that need low latencies. Also, mobile edge computing (MEC) allows for moving the application closer to the end device (at the edge of the mobile network), thereby eliminating the core network part of the latency for differentiated traffic. There is also a need to revisit the design of transport protocols for such MEC traffic, which can be better optimized for low latency services.

4.2.3 Testbeds and Demonstrators

An object of discussion was the capabilities that researchers would like to have on a testbed to test low latency applications and the underlying protocols needed to support them. Some of the open questions addressed were:

- Physical Layer: Do experimenters need to innovate on the PHY layer or use existing and emerging solutions? WiFi, 4GLTE and the emerging 5G bands were some of the candidates. These would cover sub- 6GHz and mmWave implementations.
- Link layer: It would be desirable to test over different environments, e.g. line of sight and non-line-of-sight, long distance suitable for macrocells down to small cell distances. It was also felt the multi-RAT implementations should be incorporated in the testbed so that sharing and aggregation across different spectrums and technologies, e.g., 5G. LTE, WiFi, etc., could be tested, given that seamless operation across them would be necessary to meet latency requirements.
- Transport Layer: TCP has severe limitations in terms of meeting tight end-to-end latency requirements. Bufferbloat, which is when buffers build up occurs when congestion or link capacities drop, lead to very large delays before TCP can adapt. Testing innovative transport layer protocols that address this issue would be important.
- Applications: Each application considered had its own distinct requirements for testing purposes. *Manufacturing* operates in a harsh environment with interference and a challenging propagation environment. Hundreds of thousands of sensors, typical of a manufacturing plant, need to be emulated or simulated to create a credible testbed. Such end device scaling issues and the power constraints of sensors can then be studied. *Virtual reality:* Measuring delay and jitter to meet the exacting delay requirements for this application needs to be instrumented. These are viewed as short range currently, but are possibly long range if multiple participants share a common environment. *Telemedicine:* There are severe restrictions on testing on humans or even lab animals, and it may not be even necessary to do this. It was suggested that capturing traces off data interactions in existing telemedicine setups and using them in testbed to test out new concepts.
- Computation, server, mobile edge computing (MEC): To adequately model the end-to-end performance of wireless networks, it is necessary to co-design the applications and the underlying protocols. To do this, access to a flexible architecture that incorporates these elements would be desirable.
- Controlled mobility needs to be introduced to test mobility's impact on applications.
- Wide area scale: Both city and inter-city scale networks should be incorporated in the testbed to enable latency measurements. Similarly, rural area scenarios need also be incorporated, since rural broadband access is an urgent national need.
- Instrumentation: Measuring sub-millisecond latency reliably will be a challenge in such testbeds, but will be crucial to the success of this endeavor. Precise synchronization across network elements may be necessary.
- More generally, an open interface for experimenters where they can access all aspects of the testbed would be a challenging task. The examples of WINLAB's Orbit and Utah's PhantomNet platforms were cited.

- Incentivizing other players to collaborate on testbeds would be key to the success of a testbed. A possible partner are the NIST-supported National Network for Manufacturing Innovation institutes. Another possibility is time sharing spectrum on a day/night basis with a cellular carrier in return for access to network infrastructure.
- Security technologies of the testbed need to be tested as well. Incentivizing the security community to participate would be crucial to devise secure solutions.

Appendix A: Workshop Participants

Suman Banerjee, University of Wisconsin-Madison

Kira Barton, University of Michigan

Randall Berry, Northestern University

Rich Brown, NSF

Sujit Dey, UCSD

Atilla Eryilmaz, Ohio State University

Andy Fan, Oregon State

Henry Fuchs, University of North Carolina

James Gigrich, Keysight

I-Hong Hou, Texas A&M University

Kyle Jamieson, Princeton

Bhashkar Krishnamachari, USC

Thomas R. Kurfess, Georgia Tech

Bill Lawton, Interdigital

Eytan Modiano , MIT

Jack Nasielski, Qualcomm

Shiv Panwar, NYU

Yury Poliyanski, MIT

Bozidar Radinovich, Microsoft

Glenn Recart, US Ignite

Dola Saha, SUNY-Albany

Ashu Sabharwal, Rice

Sanjay Shakkottai, UT-Austin

Peter Steenkiste, CMU

David Strobinski Boston University

Karthik Sundaresan

Mike Wittie, Montana State University

Lei Ying, ASU

Hongwei Zhang, Wayne State University

Xinyu Zhang - U. Wisc

Wenyi Zhao, DAQRI

Gil Zussman - Columbia University

Appendix B: Workshop Agenda

Thursday, November 3:

8:00 Breakfast

8:30 opening remarks -Thyaga & Eytan

8:45 – 10:00 Short presentations

Henry Fuchs, VR and telemedicine

Thomas R. Kurfess, Gtech (automated manufacturing)
Wenyi Zhao, DAQRI, (Virtual Reality)
Glenn Recart US-Ignite
Kira Barton, UMich
David Strobinsk, BU

10:00 – 10:30 Break

10:30 – 11:30 Panel discussion - Q&A on requirements

Panelists:

Henry Fuchs, VR and telemedicine
Thomas R. Kurfess, Gtech (automated manufacturing)
Wenyi Zhao, DAQRI, (Virtual Reality)
Glenn Recart US-Ignite

Moderator: Kyle Jamieson
Scribe: David Strobinsk

11:30 - 12:30 Breakout session

Establish requirements for application domain

3 groups: Telemedicine, VR, manufacturing

12:30 – 1:30 Lunch

1:30 – 2:00 Report back from breakout sessions

2:00 – 3:30 Breakout sessions - strawman architecture

3:30 - 4:00 Break

4:00 - 5:00 Reportback and feedback from application domain

Friday, November 4

8:00 Breakfast

8:30 – 9:30 Talks on Technology limitations and capabilities

Jack Nasielski, Qualcomm)
Bill Lawton, Interdigital
Shiv Panwar, NYU

9:30 - 10:00 Break

10:00 – 11:30 Breakout session

What technology advances/breakthroughs are needed?

11:30 Report plan and writing assignment

12:00 Lunch

1:00 – 3:00 Breakout session

3:00 - 5:00 Small group to assemble draft report