Workshop Report

NSF Workshop on Ultra-Low Latency Wireless Networks

March 26-27, 2015

Report from the National Science Foundation funded workshop held March 26-27, 2015 in Phoenix, Arizona 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 Phoenix, Arizona on March 26-27, 2015.

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 delay. It is well-known and widely observed that the delays 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. Even common applications, such as real-time video-conferencing, suffer from poor performance due to the unpredictable nature of these delays.

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 delay performance. However, in a range of domains, a new 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 like online gaming and augmented reality that also require latency much smaller than what is achievable in today's wireless systems. Furthermore, some of these emerging applications often require deployment at a larger scale, making it even more challenging to provide ultra-low latency over the network. Section 4 of the report discusses emerging applications and their latency requirements.

In order to facilitate the emergence of such "mission critical" low-latency application, it is imperative that the research community develops new mechanisms for enabling consistently low-latency wireless networks. Technical challenges in achieving end-to-end low-latency exist at various levels. Section 2 of the report outlines both challenges as well as opportunities that need to be addressed in various domains, including network control algorithms for reducing latency (e.g., scheduling, routing), and Information Theoretic methods for improving delays at the physical and MAC layers. At the chip level, bottlenecks for designing ultra low-latency wireless NoCs are bandwidth and resource limitations of the physical layer (e.g., device technologies, wireless transceivers, on-chip routers, and antennas). Finally, much of the delay in today's networks arises at the "system" level (e.g., network architecture, protocol stack, security mechanisms, session management issues, etc.). Section 2.4 discusses challenges and opportunities for reducing delays at the system level.

In Section 3 the report outlines opportunities across disciplinary boundaries, and the potential to leverage research in different domains for design of low latency wireless networks. We note that there is a fair degree of similarity between wireless systems and wireless on-chip networks and concepts in network control and information theory that have been developed in the wireless systems context can be applied to on-chip networks.

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 delay 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.

Summary of Key Findings

The issue of unpredictable and often high latencies precludes wireless networks from being used in mission-critical environments. Even common applications, such as real-time video-conferencing, suffer from poor performance due to the unpredictable nature of these delays. Thus, improving latency in wireless networks is critical for enabling emerging mission critical applications that depend on consistent low latency. Moreover, opportunities exist for delay reduction at various levels of the protocol stack. In order to facilitate the emergence of such “mission critical” low-latency application, it is imperative that the research community develops new mechanisms for enabling consistently low-latency wireless networks. Technical challenges in achieving end-to-end low-latency exist at various levels, including: Network control, Information Theory, Networks on Chip (NoC) and the system level.

In the area of network control, there has been tremendous progress in the design of wireless network control algorithms and their performance analysis over the last decade. These methods have both fundamentally advanced the theory and impacted the way wireless networks are architected and implemented in practice. An important idea that has been established is the use of queue lengths for channel resource allocation. This has been shown to be “universally” useful in a variety of contexts, ranging from downlink scheduling in cellular (LTE) systems, to multi-hop resource allocation. Queue length based methods have proved to be invaluable for ensuring network stability, providing user performance guarantees (low packet delay) and in end-user utility maximization. A related, simple but fundamental, idea is that relating to the use of network state for network control. Indeed these algorithmic ideas have had a major impact on practice – wireless base stations today incorporate queue-based myopic control as a default option in their implementations.

In the context of **network control**, we identify a number of opportunities and research directions that present themselves in this research space. On the one hand, they call for the expansion of the state-of-art theoretical foundations to accommodate ultra-low delay demands of emerging applications. On the other hand, they address the development of scalable and efficient new algorithms that can provide the required low-latency guarantees despite the dynamic and limited nature of information that they need to work with. On the theoretical front, achieving ultra-low latency services, especially in large scale and mobile networks, calls for new foundations that explicitly incorporate the collection, shaping, and utilization of state and control information into the network controller design. In particular, this framework needs to expand the current state-of-art stochastic optimization and Lyapunov-based strategies to include the freshness, quality, cost, and value of information that guides the controller action. In addition, a broad number of research directions open up on the algorithmic front towards the realization of network algorithms at scale for ultra-low delay services. These algorithms are expected to provide optimal or close-to-optimal theoretical performance but with the low-complexity, low-overhead, and favorable scalability characteristics.

Information theory has contributed greatly to the design of modern wireless networks. Information theoretic capacity results provide sharp delineations between the reliable throughputs that are achievable in communication systems and those that are not, providing the gold standard to which new technologies are compared. Understanding of these limits has led to techniques that have had great impact on wireless networks including both cellular and WiFi systems. Some recent examples include the use of low-density parity check (LDPC) codes, the utilization of multiple antennas for both diversity and spatial multiplexing, and the use of opportunistic scheduling. It is well known that the sharpness of information theoretic capacity analysis is due in part from focusing on the asymptotic regime where code-word sizes (and thus the corresponding delays) become large, enabling one to sufficiently average over the randomness in the channel. Such results have been relevant for the design of modern wireless systems in part because the design of such systems has focused on increasing throughput (as opposed to minimizing delays) and these systems have sent data over time-slots that enable the use of "long enough" code-words. There have been some attempts at addressing delay considerations including the analysis error exponents, the study of finite-block length performance, as well as the notion of delay limited capacity and various efforts that combine queuing with information theory. However, a complete understanding of these issues is yet to emerge.

There are a number of significant challenges that must be overcome in applying information theory to the design of ultra low latency wireless networks. A long-standing challenge is to completely understand the non-asymptotic trade-offs between delays, throughput and reliability in wireless networks including both coding delays and queuing delays. More work is needed to better understand the fundamental trade-offs between coding delay/error probability and throughput in the low latency regime. Another opportunity is in exploring the robustness of information theoretic results in the regime of low latency. There are also opportunities for better understanding the role of user coordination under delay constraints. Communication models with various levels of user coordination can be studied, which will help us to understand the tradeoff and consequently the optimal schemes for distributed coordination. Finally, work on quantifying fundamental limits on network control information is needed. Open questions include: What information should be used as network state? How fast should the network state be updated? What is the impact of message passing complexity on effective throughput and delay performance?

During the last decade, we have witnessed a major transition from computation- to communication-centric design of integrated circuits and systems. In particular, the **network-on-chip (NoC)** approach has emerged as the major design paradigm for multicore systems-on-chip (SoC). The goal of on-chip communication system design is to transmit data with low latencies and high throughput using the least possible power and resources. The major challenges in traditional wire-based NoCs are the high latency and power consumption of the multi-hop links. By inserting single-hop long-range wireless links in place of multi-hop wired links, the overall system performance can be significantly improved. We should adopt novel architectures inspired by complex network theory in conjunction with the on-chip wireless links to design high-performance multi-core chips.

Bottlenecks for designing ultra low-latency wireless NoCs arise from the bandwidth and resource limitations of the physical layer. The key physical layer components include device technologies, wireless transceivers, on-chip routers, and antennas. Wireless on-chip communication can break the inherent limitations of locality and make remote cores appear closer to each other. In particular, effective use of on-chip wireless communication can lead to a breakthrough in the performance and energy of large scale platforms used in high-end servers, high-performance computing chips, and GPGPUs. As a result, wireless links can become a true enabler for complex applications and workloads that do not exhibit locality. Besides improving the data communication, on-chip wireless communication can also become an enabler for centralized and distributed control, and dynamic adaptability. Long-range shortcuts enabled with wireless links can significantly reduce the time to synchronize hundreds of cores while the whole chip changes its power state. Finally, there is a fair degree of similarity between wireless systems and wireless on-chip networks and concepts in network control and information theory that have been developed in the wireless systems context can be applied to on-chip networks.

At the **system level**, current wireless systems range from short distances (e.g. Bluetooth), indoor ranges (e.g. WiFi), wide-area (e.g. cellular) and all the way to really long distances (e.g. satellite communications). The latency achieved by current state-of-the-art systems varies significantly as a function of many key parameters, e.g. end-to-end distance between the transmitters and receivers, coverage and capacity of the network, mobility, network architecture and number of users simultaneously accessing the network. The general trend over the last two decades has been a gradual decrease in latency, both in wireline and wireless networks. For example, latency in 1G wireless was of the order of 500-1000 ms to 50-100ms in today's state-of-the-art 4G wireless networks. Ultra-low latency applications span across multiple domains and different spatial scales. For example, factory floor control applications may have all end-points on the same local area network. On the other hand, applications like vehicular control could involve end-points that span a wide area network. In general, the challenges lie in the multi-dimensional space of tradeoff between latency, spectral efficiency, cost, reliability, security and energy.

For applications contained within a local area network, the wireless media access is the major contributor of the latency. Reducing latency in local area networks will require innovations in reducing the duration of time between the need for access (i.e., when the bit is generated by the application) to the first transmission opportunity available to the node to send the bit. For the applications that span a wide area network over large spatial scales, additional delays are incurred due to intermediate data center/cloud. In this case, the overall latency is dictated by not just by the wireless access hop, but also the backhaul, the wireless core network, and data center/cloud latency. For mobile users, there is also some latency involved in locating the mobile user in the network, which is typically accomplished via the paging procedure. Reducing the overall application latency category needs a holistic approach that spans not just the wireless access optimizations, but also wireless core and cloud architecture. Finally, there are regulatory challenges with respect to implementing some of the solutions for differentially treating traffic from different applications. Such legal aspects are not in the scope of research initiatives, but should be kept in mind when devising and deploying real solutions.

In summary, today's communication networks and the Internet are largely geared towards moving latency tolerant content. This has been driven by the fact that networks to date have been largely focused on personal communications. In order to handle increasing traffic load, existing wireless networks have been designed and planned with capacity and coverage in mind. The latency implications of the different applications have mostly been an after-thought. However, new kinds of socially useful applications that require ultra-low latency are emerging in a range of critical domains (e.g., vehicular networks, smart grid, distributed robotics, and other cyber-physical systems). These applications may require two or three orders of magnitude improvement in latency over the current state-of-the-art. There was general consensus at the workshop that improving latency in wireless networks is critical for enabling emerging critical applications that depend on consistent low latency.

Thus, now is an opportune time to invest in research toward the design of low-latency wireless networks.

Table of Contents

1. Introduction and Overview

The workshop on Ultra-Low-Latency Wireless Networks was held on March 26-27, 2015 at Arizona State University, Phoenix, Arizona. The workshop's attendees included over 40 participants from Academia, Government and Industry, representing a range of perspectives, including: wireless network control algorithms, information theory, wireless network systems, on-chip wireless network. Participants were selected by invitation as well as members of the community selected through this open call for participation. The workshop is organized by Dr. Supratim Deb (Alcatel Lucent), Prof. Eytan Modiano (MIT), and Prof. Partha Pande (Washington State University), under the sponsorship of the National Science Foundation (NSF). This report summarizes the finding of the workshop, and makes recommendations for future research directions in this area.

Wireless networks have become a ubiquitous part of everyday life all around the world, and will continue to become more involved in all aspects of the living world as time goes by. Examples of these are satellite networks, cellular networks, WiFi-based local area networks, body-area networks, sensor networks and on-chip wireless networks. Each generation sees increases in throughput and spectral efficiencies over previous generations of wireless networks, and this trend is expected to continue as well. Yet, wireless networks are very unpredictable on one critical aspect: communication delay. It is well-known and widely observed that the delays 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. Even common applications, such as real-time video-conferencing, suffer from poor performance due to the unpredictable nature of these delays.

The domain of wireless network can be broadly classified in to two groups. The first one is the macro scale wireless network and the second one is the emerging paradigm of wireless Network-on-Chip (WiNoC) that used as a communication infrastructure for multicore chips. Despite recent advances, both these wireless networks are very unpredictable in one critical aspect: communication delay. It is well-known and widely observed that the delays incurred in accessing a wireless network, regardless of which flavor we choose, can vary widely. In a cellular network, for example, access latencies can vary between 1 millisecond and 100 milliseconds per packet. Home wireless networks and wireless hotspots based on WiFi (IEEE 802.11 standards) can experience one-way latencies between 1 millisecond and 300 milliseconds, and these can happen during the duration of a single web-browsing session. The issue of unpredictable and often high latencies precludes wireless networks from being used in mission-critical environments. Even common applications, such as real-time video-conferencing, suffer from poor performance due to the unpredictable nature of these delays. Similarly, in a WiNoC it is critical to exchange information between computing nodes within a certain delay to avoid unnecessary execution time penalty. Additionally, delay uncertainty in WiNoC will also lead to additional power consumption, which is a serious limitation of current multicore platforms.

Given the huge leaps made in wireless communications over the past decade, the goal of the workshop was to revisit foundational principles, identify gaps in existing architecture, design and practice of wireless networking systems to enable consistent low-latency wireless networking, and make a recommendation of future research directions for low-latency wireless networking.

2. Domain Specific Challenges and Opportunities

During the first morning of the workshop, participants were divided into four disciplinary groups (Network control, Information Theory, On-Chip Networks, and Wireless Systems) to address challenges and opportunities within each domain area. Below we summarize the findings at the disciplinary level.

2.1 Network Control

2.1.1 State of the Art in Network Control

There has been tremendous progress in the design of wireless network control algorithms and their performance analysis over the last decade. These methods have both fundamentally advanced the theory and impacted the way wireless networks are architected and implemented in practice.

There are two canonical settings for wireless access: (i) Cellular networks on licensed spectrum, consisting of large cells (kilometer-scale) with tens of mobile users accessing each base-station, and with a strict time-scale separation between mobile-to-cell association (mobility timescale) and the channel resource allocation timescale; and (ii) WiFi networks operating over unlicensed spectrum, with distributed channel access using CSMA-based algorithms.

From an algorithmic perspective, an important idea that has been established is the use of queue lengths (and variants such as HOL delay) the key weighting function for channel resource allocation. This has been shown to be “universally” useful in a variety of contexts, ranging from downlink scheduling in cellular (LTE) systems, to multi-hop resource allocation, to distributed channel allocation for spatial resource allocation (e.g., in CSMA-based systems). Queue length based methods have proved to be invaluable for ensuring network stability, providing user performance guarantees (low packet delay) and in end-user utility maximization. A related, simple but fundamental, idea is that relating to the use of network state. While network algorithms could, in principle, use he entire past history, it has been shown that greedy or myopic algorithms that periodically solve optimization problems whose parameters depend only on instantaneous state (current channel, queues, and topology state) lead to great network optimality properties. Indeed these algorithmic ideas have had a major impact on practice – wireless base stations today incorporate queue-based myopic control as a default option in their implementations.

From a performance analysis perspective, several new advancements have proved to be powerful for both design and analysis. Tools emerging from the merger of optimization and stochastic methods have especially been fruitful, with the methods suggesting new queueing structures for tracking and combining multiple (packet and non packet) objectives. These techniques (including Lyapunov analysis) for wireless networks have pointed toward the optimal network architectures and algorithm structures, bounds for various performance objectives, and a nuanced understanding of the trade-offs between multiple objectives. Probabilistic methods (including large deviations and mean delay analysis) have been successfully developed for sharply characterizing performance in various scaling regimes. Finally, rigorous methods inspired by statistical physics have enabled the analysis of spatially and temporally dynamic interactions resulting from the coupled air interface in distributed allocation settings.

In summary, the last decade has been an exciting time for designing and analyzing control algorithms for wireless networks.

2.1.2 Bottlenecks and Challenges

As we move forward to enable an ultra-low-latency network, it is important to realize that the wireless setting will evolve as well. Applications requiring ultra-low latency typically have features that are significantly different from existing ones. The two distinct settings of cellular networks and random access networks described earlier will likely merge into a continuum to one consisting of many small cells with fewer users per "cell" and faster mobility between cells. These cells with heterogeneous capabilities will support a mixture of cellular and device-to-device traffic over shared spectrum. There will also be a tremendous increase in available system bandwidth, with resource allocation occurring over much smaller time-scales. Importantly, we will need to re-examine the existing time-scale separation between channel resource allocation and user mobility will likely disappear.

With this new setting and the more stringent performance requirements, several new challenges emerge as outlined below.

2.1.2.1 ***Algorithmic Paradigms in the Ultra-low Latency Regime:*** As discussed earlier, network resource allocation decisions depend crucially on the current queue lengths and the instantaneous channel and topology state. However, as we move to the ultra-low latency regime, it is likely that typical queue lengths for ultra-low latency applications would be very small. Further, many emerging applications have different features from traditional ones. For example, applications may drop packets that exceed deadlines, and other applications might focus on the delay of information instead of the delay of packets. Thus, queue lengths may not be fully informative for network resource allocation and decision-making. New resource allocation paradigms that inclusively take into account all indicators of performance for the ultra-low delay regime are needed.

2.1.2.2 ***Information State at Scale.*** The increased bandwidths (potentially several orders of magnitude), rapid flux in cell-to-user association, traffic heterogeneities and shorter time-scales will naturally lead to far greater amounts of state (e.g. channel, topology, neighborhoods) that needs to be measured/learned. Importantly, with increased scale (density and bandwidth), it seems likely that the amount of state grows super-linearly. The implication is clear – acquiring and decision-making based on complete state is infeasible. Thus, one needs to move away form a "complete network knowledge" setting that is dominant today, to one where resource allocation occurs with partial and noisy estimates of network state.

2.1.2.3 ***Rethinking the Control Plane.*** As we move towards the ultra-low latency regime, the overheads of protocols, often ignored in network control formulations, become key bottlenecks. For instance, current architectures incur a very high latency, on the order of tens to hundreds of milliseconds, for identity authentication, handshaking, handoff, requesting and granting transmission opportunities, retransmissions, and switching between sleep/active states. These overheads, currently being amortized due to existing time-scale separations between resource allocation and user dynamics, need to be factored into future architectures. A principled and comprehensive theory for the wireless control plane is sorely lacking. This theory needs to address various control plane design choices (e.g. mode of handshaking, open-loop vs. closed-loop connection setup, air interface access mechanisms for the control channel), and characterize their fundamental limits and trade-offs.

2.1.2.4 ***Multi-Hop Coordination.*** With the expected densification and heterogeneity of future networks, wireless communications will likely involve more than one hop of relaying. Indeed in this setting, many access points might themselves rely on wireless backhauls to relay traffic amongst them to reach a wired connection to the Internet. Multi-hop coordination inherently leads to an increase in uncertainty in several dimensions – signaling, channel access, queueing, mobility, etc. – each of which introduces additional variability in latency.

2.1.3 Opportunities and Future Directions

In view of the above bottlenecks and challenges towards the realization of ultra-low latency network control, we identify a number of opportunities and research directions that present themselves in this new research space. On the one hand, they call for the expansion of the state-of-art theoretical foundations to accommodate ultra-low delay demands of emerging applications. On the other hand, they address the development of scalable and efficient new algorithms that can provide the required low-latency guarantees despite the dynamic and limited nature of information that they need to work with. Next, we discuss some of these opportunities and directions in detail, together with specific technologies and solution strategies that may be useful in their resolution.

On the *theoretical front*, achieving ultra-low latency services, especially in large scale and mobile networks, calls for *new foundations that explicitly incorporate the collection, shaping, and utilization of state and control information into the network controller design*. In particular, this framework needs to expand the current state-of-art stochastic optimization and Lyapunov-based strategies to include the *freshness, quality, cost, and value* of information that guides the controller action. A deeper understanding of information on the design and operation of network controllers is critical in advising the nature and amount of *limited and partial information* that can suffice to achieve the latency demands of large-scale mobile networks with *small overhead.*

Another open research direction that pertains to theoretical foundations concerns the derivation of *fundamental bounds on the achievable low-latency regions.* The nature of latency-related metrics required by different application can be quite diverse, ranging from strict deadline constraints and regular service guarantees to large-delay bounds. In view of the heterogeneity of these demands, there is a need to characterize the region of achievable latency-metric as function of traffic statistics and requirements. These characterizations will not only work as benchmarks to aspire to, but can also provide guide the design of controllers that maximize latency-constrained services.

In addition to the above, a broad number of research directions open up on the *algorithmic front* towards the realization of *network algorithms at scale for ultra-low delay services.* These algorithms are expected to provide optimal or close-to-optimal theoretical performance but with the low-complexity, low-overhead, and favorable scalability characteristics. The ultra-low nature of latency demands together with the dynamic and large-scale nature of underlying networks require exploration of new paradigms than those prominent in today's network algorithm designs. Some of the promising paradigms are outlined next.

In view of the stringent latency requirements and dynamically changing network state information, one direction of research is the design of ultra-low latency Multiple-Access-Control (MAC) schemes that avoids the cost of collision resolution protocols, such as CSMA-variants, that exhibit large delay tails in moderate to heavy loaded traffic conditions. One research in this direction in dynamic and dense conditions is the development of *delay-free MAC* solutions whereby ultra-low latency demands access the common medium as soon as necessary with reliability requirements incorporated into the PHY layer. Another promising research direction is noted to be the use of *randomized algorithms* that can harness the averaging effect in large-scale networks without the heavy overhead costs of information exchange.

Another exciting research area that will be a key enabler for ultra-low latency network services is the combination of *statistical information* with the *observed state information* for control. One promising direction in this space is the collection and use of historical data at wireless *data centers* to learn and estimate the relevant traffic and network states from limited information to better guide latency-sensitive services. Another promising direction is the development of *caching* and *content sharing strategies* that utilizes *statistical predictions of future demands* and *side information from socially connected neighbors* to push the content closer to the users before the time of actual interest, and do so in a user-privacy preserving manner. The big benefit is that these strategies have the potential to cut down the latency levels significantly, and therefore fit well with the targeted ultra-low latency services.

2.2 Information Theory

2.2.1 State of the art

Information theory has contributed greatly to the design of modern wireless networks. Information theoretic capacity results provide sharp delineations between the reliable throughputs that are achievable in communication systems and those that are not, providing the gold standard to which new technologies are compared. Understanding of these limits has led to techniques that have had great impact on wireless networks including both cellular and WiFi systems. Some recent examples include the use of low-density parity check (LDPC) codes, the utilization of multiple antennas for both diversity and spatial multiplexing, and the use of opportunistic scheduling. Information theoretic analysis of fading multipath wireless channels helped lay the intellectual foundations for Orthogonal Frequency Division Multiple Access (OFDMA), the air interface technology of nearly every modern wireless system. Information theoretic ideas such as network coding have led to new approaches for developing codes for storage, which is actively being considered to provide efficient caching in wireless (and wire-line) networks. Moreover, and perhaps more importantly, information theory has helped to guide the architecture of modern wireless networks, by justifying the use of bits as a universal currency for information. This has been provided by the so-called separation theorems that assert that asymptotically, there is no loss of end-to-end application performance in communication settings if we restrict ourselves to digital communication architectures. This established bit-rate and spectral efficiency as fundamental parameters of interest that system builders could then optimize for.

There has also been a steady-stream of important information theoretic results that extend well beyond the current state-of-the art in practical wireless networks but will provide the intellectual foundation for their future evolutions. This includes significant progress in understanding challenging problems in multi-user information theory including the MIMO broadcast channel, interference channels, two-way channels (including considerations of full-duplex communication) and various relay channels. Intriguing techniques such as interference alignment, dirty-paper coding, and cooperative relaying have emerged from this endeavor. In addition, information theorists have developed models that give insights into the scaling limits of large multi-hop networks and have also enriched their models to consider issues such as the role of feedback, the addition of security constraints, and the use of energy harvesting.

It is well known that the sharpness of information theoretic capacity analysis is due in part from focusing on the asymptotic regime where code-word sizes (and thus the corresponding delays) become large, enabling one to sufficiently average over the randomness in the channel. Such results have been relevant for the design of modern wireless systems in part because the design of such systems has focused on increasing throughput (as opposed to minimizing delays) and these systems have sent data over time-slots that enable the use of “long enough” code-words. There have been some attempts at addressing delay considerations including the analysis error exponents, the study of finite-block length performance, as well as the notion of delay limited capacity and various efforts that combine queueing with information theory. However, a complete understanding of these issues has yet to emerge.

2.2.2 Bottlenecks and Challenges

There are a number of significant challenges that must be overcome in applying information theory to the design of ultra low latency wireless networks. A long-standing challenge is to completely understand the non-asymptotic trade-offs between delays, throughput and reliability in wireless networks including both coding delays and queuing delays. For coding delays, the notion of error exponents (reliability functions) provides some insights, by characterizing the exponential rates at which error probabilities decay as coding block-lengths become large. However, this approach does not capture sub-exponential terms that are needed to get a full picture of low delay performance. Recent work on finite-block length analysis and channel dispersion helps in this regard but has not yet been developed in the generality needed by multiuser wireless networks. Furthermore, this approach is based primarily on block codes. For low latency settings, other coding approaches such as streaming codes or techniques that incorporate feedback are relevant and may offer better delay performance, but again our understanding of both codes and the fundamental performance trade-offs is limited.

Much of the current understanding in information theory is all relative to nominal models that are specified exactly in terms of conditional probability distributions and interaction topologies. Such assumptions can be justified in current wireless systems in part because the overhead needed to gain such information can be amortized over the time needed to send data and also because “higher level” techniques such as ARQ can be used to provide additional robustness to any errors. However, these justifications will no longer hold in ultra low latency networks, requiring one to more directly address robustness. There are a number of established approaches for addressing robustness in information theory including compound channels, mismatched decoding, arbitrarily varying channels and individual sequences. However, such approaches have not been fully explored in the context of low latency wireless networks. Consequently, we often have no idea how robust relevant results are. For some examples, like diversity-multiplexing tradeoffs, we know that the results are actually quite fragile. Even small perturbations of the nominal model (e.g. introducing an atom) causes results to change radically. This poses a challenge for the practical applicability of these results in the low-latency context especially because low-latency is often demanded together with high-reliability. High-reliability in communication is often about trusting the consequences more than the model.

In addition to the overhead needed to learn the underlying channel models, wireless networks also require overhead for coordinating the actions of different users. Such coordination costs are not well understood and are often ignored in information theoretic analysis, e.g. in many multiuser information theory models all users jointly select their channel codes. Again, this can be justified to some degree if latencies are long enough to amortize any coordination information, but this will not be the case in ultra low latency networks. Furthermore, wireless networks operate in inherently stochastic environments, e.g. due to bursty user traffic and mobility. This necessitates that users' actions are dynamically controlled at multiple levels of the protocol stack. Such control hinges heavily on state information exchange and is intimately tied with the complexity of control signals. By and large, fundamental limits on the amount and form of such control information are not well understood.

From the perspective of low-latency communication, the architectural insights of information theory vis-à-vis separation theorems are also problematic. This is because they again involve taking the limit of long block-lengths or large delays. Consequently, there is a real challenge in terms of understanding what low latency communication architectures should be. To resolve this, we need to have new separation theorems (or approximate separation-theorems) that tell us what "low latency" really should mean. After all, while in human-to-human communication we could imagine that the need for low-latency is an experimental observation; there is nothing like that for machine-to-machine communication. In such settings, the real underlying performance objective is something like stabilization, control performance, system responsiveness, etc. Low-latency is simply one way to help engineer systems that deliver the underlying objective. But perhaps, it is not low latency per-se that is needed. For example, recent work at the intersection of information theory and control has identified that, sometimes, it is the predictability of the delay that is more important than the actual mean value of the delay.

Much information theoretic work on wireless networks has focused on wirelessly transmitting information. As such it does not directly account for the promising benefits of exploiting storage resources for reducing latency. For instance, if frequently requested content is strategically placed at caches close to the requesters, user delays are much reduced. Furthermore, the pervasive use of caching can provide data packets locally when packet losses occur due to channel outages, thereby significantly increasing transmission reliability and decreasing delays. While caching and storage have long been active research areas, much of the existing work focuses on static and centralized settings, and is carried out in isolation from research on other network functionalities. A complete architectural view of wireless network performance with caching is needed.

2.2.3 Opportunities and Future Directions

More work is needed to better understand the fundamental trade-offs between coding delay/error probability and throughput in the low latency regime. Such work can build on work such as that on channel dispersion and the study of streaming codes with feedback and provide new insights into the design of low latency networks. Accounting for issues such as random traffic arrivals and the multi-user nature of wireless networks in such a framework are also intriguing future directions.

Another opportunity is in exploring the robustness of information theoretic results in the regime of low latency. Again there is a foundation to build on including the ideas of mismatched decoding and compound channel models, but these ideas need to be more fully developed for multiuser wireless networks with low delay.

There are also opportunities for better understanding the role of user coordination under delay constraints. Communication models with various levels of user coordination can be studied, which will help us to understand the tradeoff and consequently the optimal schemes for distributed coordination. Understanding is needed on the impact of distributed coordination in the context of network architecture in the sense that which coordination should be considered as physical layer issue, which coordination should be considered as link layer issue, and whether a specific classification can lead to fundamental architectural inefficiency in practical applications.

Also, work on quantifying fundamental limits on network control information is needed. Open questions include: What information should be used as network state? How fast should the network state be updated? What is the impact of message passing complexity on effective throughput and delay performance?

For low latency applications, it is vital to understand what the relevant system parameters actually are so that engineers can optimize for them. From an information-theoretic point of view, this calls for new separation theorems and approximate separation theorems. At the experimental level, there might also be a need for a much richer vocabulary of "latency" related concepts to properly express a more nuanced mixture of needs that even human-interacting applications require in terms of date-rate/reliability/latency.

With regard to caching, there are promising directions to be explored which investigate the joint design of caching with other network control operations such as routing, scheduling, congestion control, and possibly network coding. The focus of the joint design should be on scalable, distributed, dynamic algorithms which optimize the use of bandwidth and storage resources in the presence of changing content, user demands, and network conditions. A comprehensive research program in the strategic use of storage in wireless environments should include the study of fundamental limits of caching, as well as the design of practical and robust algorithms.

2.3 On-Chip Network

2.3.1 On-chip Communication via NoCs

During the last decade, we have witnessed a major transition from computation- to communication-based design of integrated circuits and systems. In particular, the network-on-chip (NoC) approach emerged as the major design paradigm for multicore systems-on-chip (SoC). Consequently, it became clear that two of the most important concepts that will drive the design in future years are low power and network design. Although for some time it was nearly impossible to understand what exactly a certain design can or cannot do for the application at hand, we can now compare and contrast various optimizations using realistic benchmarks and understand what benefits a network-based approach can bring to various applications. Major driver application domains of NoCs are multimedia, embedded systems, high-performance computing, general-purpose computing on graphics processing units (GPGPU), data centers, neuromorphic processors, and networking. Hence, NoC research considers both intra-chip and inter-chip communication, which encompasses a wide spectrum ranging from multi-core chips all the way to data-centers.

To give a sense of the complexity of network-based design, the NoC platforms are expected to sustain the communication among hundreds to thousands of heterogeneous cores. Platform heterogeneity is across the board from architecture to power profiles. Consequently, the NoC-based infrastructure should enable communication among heterogeneous computational nodes ranging from multiple general-purpose cores, to GPU cores and DSPs, to application specific accelerators, all with very different performance and power profiles. Moreover, heterogeneity has recently been enriched with the integration of big (high performance) and little (energy efficient) CPU cores to the same chip. Therefore, the overall power consumption of the NoC-based platform ranges from mW in low power application-specific SoCs, to few hundred of Watts in high-end servers and high performance computing applications. Likewise, the average packet latency in NoCs varies from a few nanoseconds to a few hundreds of nanoseconds.

Most common routing algorithms used in NoCs are deterministic in nature due to their simplicity. At the same time, there are applications where lightweight adaptive and stochastic routing is needed for dealing with fault tolerance and performance issues. Regardless of the choice of the routing algorithm and application domain, freedom from deadlocks and livelocks is a fundamental requirement, since they can easily paralyze the system. Similarly, wormhole routing is the most common switching technique due to its low buffer requirements (typically a few Kbytes), which is critical to minimize the area overhead of NoCs. However, circuit switching is also employed to provide guaranteed services. We also note that there are bufferless solutions that can reduce the buffer area aggressively at the expense of performance and power consumption.

Power consumption and energy efficiency have been among the leading considerations in NoC design. Low power design techniques all the way from circuit to architecture and system-level have been a hot research topic. Using multiple voltage and frequency islands (VFIs), where different regions of the NoC can run at different frequencies, has emerged as an effective knob to drive power consumption down. In particular, dynamic voltage-frequency scaling (DVFS), as well as the setting of unused resources to sleep states, have both proven to be effective to improve the overall system energy efficiency. Consequently, state-of-art NoCs have tens of different voltage islands.

Allowing multiple VFIs, hence multiple clock domains, opens up the fundamental question about the choice of communication protocols. On the one hand, synchronous communication that requires a global clock does not scale to large NoCs. On the other hand, synchronous communication is not an option when there are multiple clock domains. Hence, globally-asynchronous locally0synchronous (GALS) communication stands out as a promising solution that naturally pairs with NoCs with multiple voltage-frequency islands, where a fully-asynchronous NoC is used to integrate VFIs. Recent industrial examples illustrating the power and scalability benefits of asynchronous NoCs include STMicroelectronics' P2012/STHORM accelerator-based reconfigurable GALS multiprocessor (2012-14), and IBM's TrueNorth fully-asynchronous neuromorphic chip (2014) with 4096 neurosynaptic cores which consumes only 70mW during real-time operation.

From a reliability point of view, NoC have very stringent demands. Packet drops are rarely accepted unless some particular application domains are considered. For most of the case studies, the bit error probability has to range between 10^{-15}~10^{-6} with a special emphasis on small values. The nature and requirements of the application can make a significant difference not only in terms of algorithms and strategies suitable for enforcing fault-tolerance, but also for deciding the best mapping, scheduling, and routing approaches. For instance, the applications do not only dictate the BER, but also the timing requirements and constraints. Similarly, the number and type of errors to be overcome, as well as the timing aspects are important for choosing the optimization strategy at all levels of the design. Depending on the application domain, the NoC design has to decide between guaranteed service or best effort type of service.

Current and emerging applications exhibit complex spatio-temporal variability, and result in highly variable network traffic, which is neither memoryless, nor stationary. Even a single application can exhibit very different (highly-variable) spatial dependency among its internal tasks as a function of the input data and the state of internal computations. In more complex scenarios, the communication workloads can also display self-similar and fractal characteristics. Another particular feature of future applications is that in order to take advantage of the concurrency the amount of traffic to be communicated between cores and even external devices will increase significantly posing serious problems on networking resources. As a result, workload analysis should not be an afterthought, but rather a first-class consideration for multiprocessor platform design.

2.3.2 Bottlenecks and Challenges

The goal of on-chip communication system design is to transmit data with low latencies and high throughput using the least possible power and resources. The major challenges in traditional wire-based NoCs are the high latency and power consumption of the multi-hop links. By inserting single-hop long-range wireless links in place of multi-hop wired links, the overall system performance can be significantly improved. We should adopt novel architectures inspired by complex network theory in conjunction with the on-chip wireless links to design high-performance multi-core chips. Between a regular, locally interconnected mesh network and a completely random Erdös-Rényi topology, there are other classes of graphs, such as small-world and scale-free graphs. Small-world graphs have a very short average path length, defined as the number of hops between any pair of nodes. The average shortest path length of small-world graphs is bounded by a polynomial in log(N), where N is the number of nodes, making them particularly interesting for efficient communication while using minimal resources. Indeed, NoCs incorporating small-world connectivity can perform significantly better than locally interconnected mesh-like networks, yet they require far fewer resources than a fully connected system. A small-world network-based architecture has many short-range (local) links, as well as a few long-range links. As long metal wires are costly (both in terms of power and latency), one should use wireless links to connect the nodes that are far apart. In practice, depending upon the available wireless resources, we can only allow a limited number of long links in the wireless NoC to be wireless, while the others would still remain wireline. This way, we can make the distant cores "socialize" with each other, and hence reduce the communication costs when running real applications.

Bottlenecks for designing ultra low-latency wireless NoCs are bandwidth and resource limitations of the physical layer. The key physical layer components include device technologies, wireless transceivers, on-chip routers, and antennas. Recent investigations have established that silicon integrated on-chip antennas operating in the mm-wave and Sub-THz range represent now a viable technology. Coupled with significant advances in mm-wave and Sub-THz transceiver design, this on-chip wireless link approach opens up new opportunities for ultra low- latency wireless NoCs. The state of the art wireless NoC designs currently use wireless links operating in the mm-wave frequency range of 10-100 GHz using existing CMOS technology. Performance of the wireless NoC can be improved if the number of non-overlapping wireless channels and their bandwidths are increased. To achieve that goal, it is necessary to extend the operating range of the on-chip mm-wave wireless channels to the Sub-THz to THz range. This will eventually alleviate physical layer limitations and significantly enhance the achievable performance of the wireless NoC.

Furthermore, power consumption and thermal hotspots impose stringent constraints on the design. These challenges are currently addressed by dynamic power management systems (PMS) that aim at the optimal tradeoff between power and latency. There are two types of PMS systems commonly used for massive multicore systems, viz., Dynamic Voltage and Frequency Scaling (DVFS) and Voltage Frequency Island (VFI). Wireless NoC can aid with the efficient implementation of both these mechanisms. By reducing the hop count between largely separated communicating cores, wireless shortcuts can carry a significant amount of the overall traffic within the network. The amount of traffic detoured in this way is substantial and the low power wireless links enable significant energy savings. However, the energy dissipation within the network is still dominated by the data traversing the wireline links. Hence, the overall energy dissipation of the wireless NoC can be improved even further if the characteristics of the wireline links are optimized. Consequently, implementing DVFS on the wireline links of a wireless NoC-enabled multicore architecture has the potential for providing even more energy savings.

In recent years, multiple VFI designs have increasingly made their way into commercial and research multicore platforms. Each VFI in wireless NoC architecture can implement a suite of power management capabilities and exploit the small-worldness via the wireless shortcuts in order to make the power management process more efficient.

The multiscale behavior of both computation and communication workloads call for a restructuring of the network design and optimization techniques irrespective of whether we consider wired communication only or mixed wired and wireless communication. The high spatio-temporal variability of emerging applications in genomics, proteomics, big data, graph analytics and decision-making also make static mapping and scheduling obsolete, since by the time a static assignment is done, the characteristics of the application may change completely. Similarly, the optimization of wireless NoC platforms has to take into account the dynamics of workloads and contribute to the creation of self-organizing architectures that can adapt them to meet the desired performance or power budget. Hence distributed and asynchronous protocols are needed to dynamically reconfigure both NoC topologies and routing policies, based on actual evolving application characteristics.

Another challenge is to share the available resources efficiently with the goal of maximizing the resource utilization and minimizing power consumption. Hence, dynamic management of resources, i.e., runtime optimization, is critical to adapt to highly varying application characteristics. One related challenge is that as NoC platforms scale in size and functionality, it becomes prohibitive to collect information about the global state or utilization level of network resources. What is more, wireless transceivers do not scale with number of cores. Inducing intelligence within the wireless NoC platforms requires both the development of mathematical techniques for analyzing randomized strategies that evolve over networked architectures and distributed algorithms that can reconfigure the architecture to meet the application requirements and power/thermal budget.

Developing a new mathematical theory of network design together with better device technology and sub-THz and THz circuits can alleviate the aforementioned challenges. However, technology alone is not sufficient to address these challenges and provide the required improvements. Better resource management, utilization of bandwidth, channel, divide between wired/wireless cross-layer innovation can lead to a few orders of magnitude reduction in latency (protocol, algorithms, architecture, circuits). On-chip wireless communication can also enable local control with full state information and global control with partial state information. Towards this end, on-chip small-world networks will help convergence and improve adaptability. The wireless links can be used not only to transport communication workloads, but also to help synchronize distant parts of the chip by providing means for fast dissemination of control strategies. This requires a new breed of tools that can assist designers to evaluate various trade-offs.

One fundamental challenge towards developing this new theory of network design is concerning the mathematical modeling of computation and communication workloads. There is a need for coupling within the mathematical models of workloads the technological constraints of wireless transceivers and determine the best tradeoffs between performance, power and design complexity. We cannot afford to profile applications and perform offline optimizations, but rather we need to endow the NoC platforms with capabilities of developing and learning the mathematical models of the workloads from real-time observations in a distributed fashion and use these as premises for further adaptation and reconfiguration of resources. Such a mathematical theory has crucial importance for power, thermal and reliability management. Investigating the tradeoffs between the complexities of distributed mathematical modeling and learning that should be enforced within the hardware and the topology used for adaptation requires fundamentally new algorithmic approaches.

Finally, security of interfacing the wireless chips with other systems and cloud is also an important challenge that needs to be addressed.

2.3.3 Opportunities and Future Directions

Wireless on-chip communication can break the inherent limitations of locality and make remote cores appear closer to each other. In particular, effective use of on-chip wireless communication can lead to a breakthrough in the performance and energy of large scale platforms used in high-end servers, high-performance computing chips, and GPGPUs. As a result, wireless links can become a true enabler for complex applications and workloads that do not exhibit locality.

Besides improving the data communication, on-chip wireless communication can also become an enabler for centralized and distributed control, and dynamic adaptability. Long-range shortcuts enabled with wireless links can significantly reduce the time to synchronize hundreds of cores while the whole chip changes its power state. At the same time, wireless links can alleviate congestion and floor planning problems, which is a daunting challenge in large chips.

Another promise of wireless on-chip communication is to improve thermal behavior and reliability. Specifically, on-chip wireless links using near field inductive coupling (NFIC) can alleviate temperature hotspots and reliability by eliminating or reducing physical TSVs in 3D ICs. Hence, contactless wireless link in 3D ICs can be a promising solution.

Last but not least, the analysis of wireless on-chip communication will also lead to new mathematical models and algorithmic developments that will benefit not only the design of wireless based NoC platforms, but can also serve as theoretical premises for developing solutions to other domains where wireless communication can be exploited such as cyber physical systems and internet of things.

2.4 Wireless Systems

2.4.1 State of the Art

Current wireless systems range from short distances (e.g. Bluetooth), indoor ranges (e.g. WiFi), wide-area (e.g. cellular) and all the way to really long distances (e.g. satellite communications). The latency achieved by current state-of-the-art systems varies significantly as a function of many key parameters, e.g. end-to-end distance between the transmitters and receivers, coverage and capacity of the network, mobility, network architecture and number of users simultaneously accessing the network. The general trend over the last two decades has been a gradual decrease in latency, both in wireline and wireless networks. For example, latency in 1G wireless was of the order of 500-1000 ms to 50-100ms in today's state-of-the-art 4G wireless networks.

To understand the components of delay in a typical large-scale network, consider the architecture of today's 4G LTE network as illustrated in Figure 1. The end-to-end delay typically comprises of three parts: the over-the air delay, the delay in the mobility/wireless core network (between the cell site and the boundary of operator network such as a P-GW, i.e., Packet Gateway), and the delay from P-GW to the application service provider in the Internet cloud. The mobility core network plays the important role of providing seamless connection to devices even as the user moves across multiple cell sites by anchoring the packet flow and tracking the user's mobility. The delay between the device and the eNodeB is driven by the air-interface design. LTE has sub-frame durations of 1 ms which is the minimum time required to send a packet. However, with a scheduling grant delay of 4ms in the uplink and the hybrid-ARQ delay of 8ms for each retransmission, the round-trip over-the-air delay typically comes to around 20ms. On the other hand, the wireless/mobility core network latency depends on the locations of the mobile gateway node and the application server. Typically, the mobile gateway node in current networks is centralized, each serving several million customers. Application servers may be hosted in data centers not necessarily co-located with the mobile gateway. In the case of content distribution, it is increasingly common to have a cache at the mobile gateway site to reduce the latency and improve the user experience. On an average, the delay experienced by a packet in the wired network ranges from 40 to 80 ms, depending on the distance from the cell site to the P-GW, the distance from the P-GW to the data center, etc.

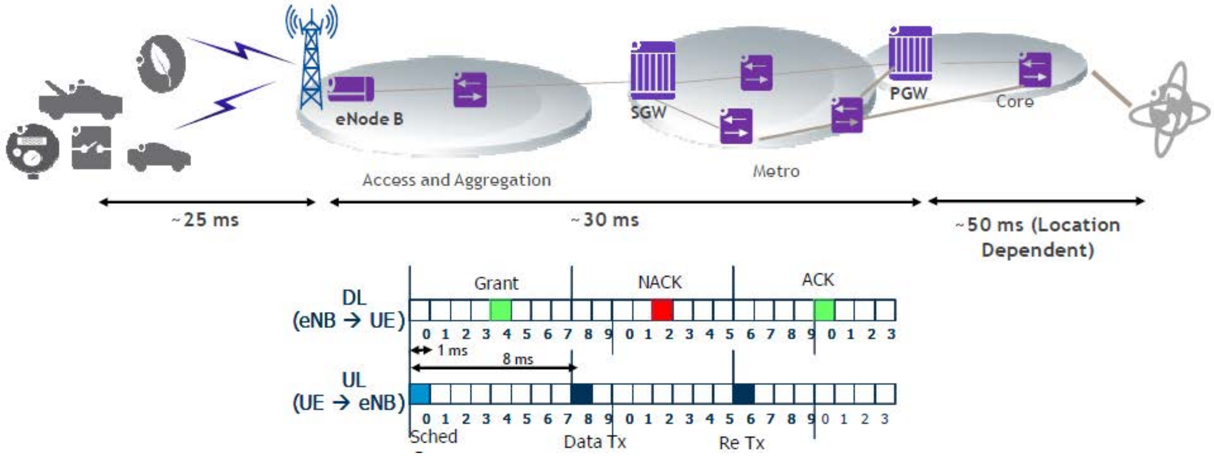


Figure 1: Architecture of 4G LTE Network, with approximate latency in different network elements. The bottom picture highlights the source of major delay in wireless access layer.

2.4.2 Bottlenecks and Challenges

The architecture example of Figure 1 and the above discussion illustrates the challenges in current large-scale networks. Ultra-low latency applications span across multiple domains and different spatial scales. For example, factory floor control applications may have all end-points on the same local area network. On the other hand, applications like vehicular control could involve end-points that span a wide area network. In general, the challenges lie in the multi-dimensional space of tradeoff between latency, spectral efficiency, cost, reliability, security and energy.

For applications contained within a local area network, the ***wireless media access*** is the major contributor of the latency. Reducing latency in local area networks will require innovations in reducing the duration of time between the need for access (i.e., when the bit is generated by the application) to the first transmission opportunity available to the node to send the bit. The access delay is highly dependent on overall capacity and the number of competing nodes on the wireless channel. Due to lack of a coherent theory of access delay in general networks (without any constraints on access technology), it is also not clear if we are operating close to the fundamental limits. Some of the specific challenges in reducing latency in the wireless access are:

1. Redesigning media access protocols so that different applications that have different latency requirements are provided appropriate latency guarantees for contention latency (for CSMA access), queueing latency, and transmission latency (via MCS/rate assignment).
2. Devising hybrid ARQ and ARQ mechanisms that use different parameters for different application types (even within the traffic of a single user) to provide the desired retransmission latency to each application/service type.
3. Devising fast beamforming algorithms that have low latency overheads during the learning phase of beamforming.
4. Designing fast encoder/decoder, scheduler, and security engines for rapid runtime framing and encryption/decryption of over-the-air packets, so that scheduling decisions can be made at a finer time scale. This amounts to reducing the MAC scheduling interval (e.g., 1ms subframe duration in LTE). With smaller scheduling timescales, the shadow-fading variations and channel prediction algorithms will need to be revisited.
5. Devising fast cooperative transmit/receive schemes to reduce latency due to joint encoding decoding to leverage multiple base stations transmit/receive a data stream to a single user without incurring increased delay due to joint operation.

For the applications that span a wide area network over large spatial scales, additional delays are incurred (as explained in Figure 1) due to intermediate data center/cloud. In this case, the overall latency is dictated by not just by the wireless access hop, but also the ***backhaul, the wireless core network, and data center/cloud latency***. For mobile users, there is also some latency involved in locating the mobile user (or equivalently, determining the cell site where the user currently resides) in the network which is typically accomplished via the paging procedure. Reducing the overall application latency category needs a holistic approach that spans not just the wireless access optimizations, but also wireless core and cloud re-architecture/rethink. Some of the specific challenges in reducing latency of the wireless core and/or cloud are:

1. Redesign the wireless core network and cloud architecture and deployment design to reduce the impact of propagation delay between the cell sites and wireless core gateways.
2. Revisit and redesign the mobility core network design and session management procedures defined in the standards to reduce the latency when user attaches to the network, as well as when the user transitions from idle to active mode.
3. Revisit the well-established paging procedures to reduce the device discovery latency. For example, an application class based paging mechanism, which searches for a device more or less aggressively (at higher or lower signaling and state management costs respectively) depending on the application traffic type. The same device could be paged more aggressively to send an important alarm message, but less aggressively to send a status update message. The challenge here is that the entire network (radio and core) needs to be application aware.
4. Reducing the propagation delay of ultra latency sensitive applications in the mobility core network by allowing the traffic to flow in a peer-to-peer manner directly from device to cell site to the destination cell site and then to the other device, instead of requiring it to first go through the wireless core network gateways.

5. Devising new deployment strategies and operator-provider business models whereby the latency between the operator core network and the provide cloud can be reduced through co-location close to the edge. For example, running a common cloud near the edge that not only hosts the wireless core gateways (SGW, PGW, etc) but also the application provider servers.

In scenarios where today's wireless networks cannot meet the latency requirements of certain applications, customized wired networks are employed. For example, an end to end latency guarantee required for a power grid monitoring system is of the order of 17ms (1/60Hz) so that a distribution center can detect fault and take immediate action to trip circuits at downstream power stations when a failure is detected. With current LTE architecture, it is extremely difficult to provide such guarantees for these applications. These and other industrial monitoring M2M networks today use optical and/or copper infrastructure for interconnecting the devices, often through dedicated optical networks. Other applications such as high-frequency trading use custom designed point-to-point microwave links. However, the ***capital and operational expenditure*** for such dedicated infrastructure is prohibitively high. The challenge is to meet the same reliability and latency guarantees that these wired subsystems deliver by using wireless networks.

Although not a technical challenge, there are regulatory challenges with respect to implementing some of the solutions for differentially treating traffic from different applications. For example, can different applications be provided different latencies based on their needs without violating net neutrality rules. Such legal aspects are not in the scope of research initiatives, but should be kept in mind when devising and deploying real solutions.

2.4.3 Opportunities and Future Directions

To support novel and emerging applications that demand ultra-low latency, it is crucial that we need to develop new foundations, practical design and experimental prototypes. Low latency is widely accepted in the wireless industry as one of the seven key "rainbow of requirements" that will enable 5G wireless communications, and has been incorporated in the ITU-R framework for defining 5G communications. Several new research directions that are crucial to achieve ultra low-latency are described below.

1) **Delay-sensitive PHY and waveform designs**: To meet the the low-latency requirements of next-generation ultra-low latency services, the traditional PHY and waveform designs need to be reconsidered to allow for more flexible scheduling with reduced latency. To this end, there are a few promising directions that are worth further exploration. First is the emergence of infrastructure nodes that can use very large number of antennas. Next generation cellular standards have already begun standardization process of 64-antenna two-dimensional arrays. The high number of antennas provides an opportunity to design low-latency waveforms. Second is the availability of large amount of bandwidths in higher frequencies, such as 28/38 GHz. Compared to current technologies working at low frequencies (sub-3 GHz), the numerology may be scaled down for higher frequencies such that individual transport blocks may be transmitted within dozens of micro-seconds, enabling air-interface round-trip delays of around 1ms. In addition, new PHY waveforms such as filtered OFDM may be developed which are more adaptive to different deployment scenarios and support the multiplexing of low-latency and normal traffic in a more robust manner.

2) **Service-aware Medium Access Protocols**: As discussed above, medium access adds the largest delay in current air interfaces. A promising avenue is design of service-aware access protocols that could provide multiple levels of latency-sensitive traffic support. The key design driver has to reduction in the control overhead that includes handshaking, security, and feedback like H-ARQ. For new applications such as V2X, radio sleep techniques should be jointly designed with medium access protocols. For example, standards frameworks such as 802.11e and QCI-based (QCI stands for *QoS Class Identifier*) scheduling in LTE need further refinement to accommodate ultra low latency applications.
3) **Analytics-driven Network Stack:** A key research direction is the ability to harness existing, and “ambient information” to improve network design. Some promising possibilities of using ambient information are using one user’s channel estimates for another user’s link and leveraging coarse location estimates. The cell sites can adopt an analytics-driven approach for detecting mobile users, their activity patterns, wireless signal signature, and their mobility patterns to pre-emptively initiate media access, authentication, and handover procedures to cut down on wireless latency. The neighboring cell sites can use collective data analytics on the behavioral patterns of a given user. Furthermore, the processing power on the devices could be used to aid such procedures.
4) **New Backhaul Architectures**: Backhaul delays are dominant contributor of latency in large-area networks. Several promising research directions have been identified. First, with processing power becoming cheaper in the datacenter, as well as on the end devices, the overall network architecture can take advantage of this to trade off computation for reducing latency. For example, the mobility core infrastructure (MME, SGW, PGW, mobile proxy, lawful intercept gateways, virus/worm detection DPI gateways, etc) can be moved closer to the cell sites. Instead of having a handful of centralized nationwide data centers where wireless traffic is funneled (thereby incurring significant propagation latency), building a large number of smaller data-centers closer to the cell sites using the ever cheaper server hardware can reduce the wireless core and cloud latency. Secondly, for long-distance backhaul links, low-orbit satellites or high-altitude balloons could serve as faster links, compared to an all-wireline backhaul. Thus, there is a need for research in designing mixed mode backhaul architectures, that leverage wireline and long-range wireless.
5) **Security as an Important Design:** Security is another important consideration that can impact latency on multiple levels. The handshaking protocol for network authentication and processing complexity for encryption/source identification needs to be revisited. As we do for increased multi-Gb/s data rates, cross-layer encryption approaches for data/control need to be investigated.
6) **Application Class-based Network Design & Slicing**: With a mix of regular user web browsing traffic and M2M/M2H traffic, it is necessary to redesign the network such that the the network can be flexibly and easily partitioned into slices to satisfy the QoS requirements of both types of traffic using a common network infrastructure. Within the M2M/M2H category, there are subcategories of applications that have different latency and reliability requirements - e.g., applications that need 1-10ms vs applications that need 10-30ms or 30-200ms.
    a) MAC scheduling and application-based ARQ over wireless hop can be used to differentiate traffic and provide the desired latency over the wireless hop.

b) Software defined networking (SDN) is a promising avenue to accomplish this network slicing over the wireless core network and the cloud portion of the network over which the application traffic traverses. Furthermore, pushing the wireless core and cloud closer to the edge (closer to the wireless cell sites) can enable reduction of latency.

c) Other architectural enhancements that can be considered, involve having the application traffic of certain applications that are extremely latency sensitive, flow directly between the cell sites without traversing the wireless core and cloud (short-circuiting). To meet the billing and regulatory requirements, 3GPP architecture can be evolved so as to allow cloning/copying of such traffic towards the core/cloud.

3. Cross-disciplinary Challenges and Opportunities

In this section we discuss challenges and opportunities across disciplinary boundaries, and the potential to leverage research in different domains and apply them to a particular domain for design of low latency wireless networks. We note that there is a fair degree of similarity between wireless systems and wireless on-chip networks and problems in network control and information theory for network control can apply to both networking scenarios.

3.1 Wireless systems and on-chip intersection

The traditional wireless communication system and wireless Network-on-Chip (WiNoC) address two different application domains. However, several opportunities for interaction, collaborations and cross-domain research between these two paradigms have been identified.

The on-chip network architecture is a hybrid wired and wireless communication network. Short distance local communication is preferably achieved over conventional wired links while long distance communication is realized over low-latency, low-power on-chip wireless interconnects. As the on-chip wireless medium is a shared resource, there’s an interest in maximizing its utilization and applying the wireless network to problems that benefit from the low latency interconnection that wireless provides.

Below we discuss opportunities for cross-fertilization of innovation between traditional wireless networks and on-chip wireless networks

3.1.1 Topology reconfiguration and control in response to changing state or environment

One of the greatest commonalities can be found between wireless on-chip networks and wireless mobile ad-hoc networks. This may seem surprising, as on-chip networks are physically static. However, task migrations to balance on-chip constraints such as heat and latency and the ability to reconfigure network topologies with wireless capabilities change that notion. Dynamically changing traffic patterns create network dynamics, which can leverage similarities with ad-hoc networks.

There are important differences between traditional wireless ad hoc networks and on-chip networks as well. One example is the routing goal being single-hop wireless links between embedded computing cores in the on-chip network to achieve ultra-low latency communication, whereas typically in ad-hoc networks communication is multi-hop. Hence, low-latency communication policies designed for on-chip environments can be reinvented to reduce latency in wireless ad hoc networks. In addition, mobile ad-hoc networks are less aware about traffic and device node locations, and do not control where wireless devices are in the physical world. In contrast there is a higher degree of controllability in on-chip networks. Hence, there are a number of innovations in mobile ad-hoc network that are applicable to on-chip networks if we turn around constraints and free variables. Mapping the problems into the on-chip environment can be expected to stimulate new research challenges stemming from radically different environments and targets. The scale of tolerable latency in on-chip wireless networks is several orders of magnitude lower than traditional wireless systems. This will require innovations in problem solving in the domains of topology reconfiguration and control with possible collaborations between experts in both communities with resulting impacts on both.

3.1.2 Workload and traffic modeling

Understanding traffic interaction between processing cores through the on-chip wireless network will help in better predicting traffic demands and consequent reorganization or reconfiguration of the networks. Traffic modeling and prediction mechanisms combined with a reconfigurable wireless network topology can provide tailor made solutions for the on the chip environment. This is coupled with the need for distributed control and negotiations for access and utilization of the on-chip wireless resources. Both on-chip wireless networks and macro scale wireless networks can be heterogeneous resulting in similar solutions for channel utilization and access.

3.1.3 Ultra low latency and strict power budgets

On-chip wireless networks operate under strict power and delay constraints. This translates to strict energy and latency constraints in communications. On-chip wireless networks are used for extremely low latency long distance direct communication between distant points on the die. Cross-layer innovations towards satisfying these strict latency and energy constraints can be leveraged for low-latency macro-scale wireless systems and vice-versa.

3.1.4 Theoretical models and bounds on achievable benefits

Theoretical bounds on latency-energy trade-offs given dynamic workloads/task models are necessary for both on-chip as well as traditional wireless systems to understand the range of possible benefits and techniques for approaching performance predicted by the bounds. The effort to characterization of the fundamental limits in the two domains can benefit from ideas generated in each.

3.1.5 Interaction and inter-network communication

Interaction and data communication between on-chip wireless networks and inter-chip data channels using similar or disparate wireless technologies need to be investigated for understanding potential benefits and risks. Inter-chip data channels may span multiple scales of distances from board level modules to full datacenters. This is similar to heterogeneous wireless cellular networks with small cells access and wide area backhaul. Questions and solutions around the use of same technology for access and backhaul are potentially applicable to on-chip and inter-chip networking.

3.2 Network theory and Network on Chip Intersection

Below we discuss opportunities for cross-fertilization at the intersection of network theory and on-chip wireless networks

3.2.1 Queuing and Buffer Management

In both on-chip and wireless networks, queueing management and buffer sizing have a critical impact on overall system latency. However, an important difference in the NoC domain is that buffer size has a significant cost impact on area, and dynamic and static power. In each case, buffers are used as a control mechanism to provide information on network state, and also to absorb stochastic traffic network uncertainty to improve performance. For hybrid wireless/wired NoC's, there are unique challenges, since non-selected queued packets must be diverted rapidly to long-range wireline channels rather than remain queued for pending wireless resources. Fundamental to both domains are the mathematical models of computing and communication workloads for effective support of heterogeneous traffic (e.g. handle mixes of low-latency vs. high-throughput traffic). Some important mathematical characteristics of workloads are the non-stationary and multi-fractal behavior. Even though some analytical studies have identified these characteristics, there are still numerous cases when particular features are not fully captured by current models. For instance, capturing nonlinearities and pseudo-periodicities in networking workloads remains still an open problem. In addition, for NoC's, there is a need to develop effective techniques to handle guaranteed service vs. best effort traffic (or combinations of them) which pose supplementary challenges for mathematical modeling. Such mathematical models are needed not only for generating realistic networking traces, but also for dynamic optimization such as communication scheduling and/or power / thermal management.

3.2.2 Developing Theory to Support Latency Combined with Other Cost Metrics

In both on-chip and wireless networks, a theoretical focus on latency in isolation is insufficient for realistic applications. Instead, theoretical techniques are need to model combined cost metrics as a function of system features (e.g., topology, buffer sizing, routing arbitration). For on-chip networks, wireless links currently combine lower bandwidth with significantly improved latency, over long-range wireline links. Hence, throughput and latency metrics, along with power and/or thermal figures must be combined. For wireless networks, latency also cannot be considered in isolation; instead, factors such as packet drop rate and age of information must also be jointly optimized. In each case, new theoretical tools are required to handle these more complex cost metrics.

3.2.3 More Realistic Latency-Oriented Traffic Models

For both on-chip and wireless networks, spatio-temporal models of traffic patterns for evaluating latency are typically quite limited. For example, for wireless, IID (independently and identically distributed) traffic models are often assumed for tractable performance analysis. For on-chip networks, simple exponential distributions are often used. For the latter domain, recent work has begun to consider more sophisticated models, capturing bursty and realistic flows analytically, such as self-similar and multi-fractal approaches. Furthermore, in each domain, there is a greater emphasis on analytical techniques for throughput, given the challenges in accurate modeling of latency. Hence, a re-orientation towards system latency is a critical direction for future network theory.

3.2.4 Physics-aware Mathematical Modeling for Performance Evaluation

Much of the performance analysis in both wired and wireless on-chip networks derive the throughput and latency metrics not only ignoring the power costs but also without accounting for interference or reliability constraints in the communication. There is a need to couple the performance analysis with specific features of the physical communication such as antenna design and communication channel properties. One important feature of the mathematical models is that they should also account for architectural features such as topology, queueing disciplines, routing delays.

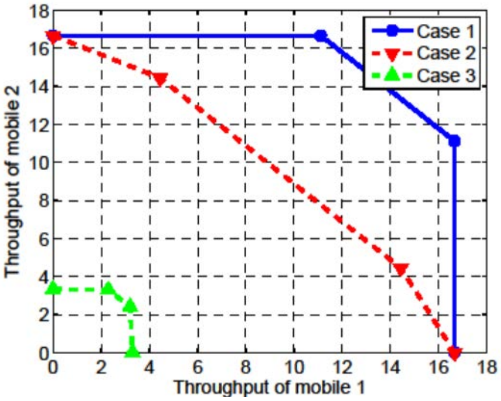
3.2.5 Traffic- and Application-aware Network Management Strategies

Many of the network optimization and control problems rely or require accurate mathematical models of the network dynamics. Power / thermal management, mapping, scheduling or routing require efficient implementation of distributed optimization and control techniques based on such accurate mathematical models. Neither centralized, nor fully distributed approaches proposed to date are likely to work in future network platforms. Consequently, the mathematical models of the workloads that can be learned at run-time not only provide information on the network state, but can also guide the hierarchical control and optimization process (e.g., topology creation and reconfiguration). Exploiting such an approach is likely to overcome the scalability issues of current approaches and contribute tremendously to scale out the multicore to the on-chip data center architectures.

3.3 Network Control and Information Theory Intersection

Next we discuss opportunities for cross-fertilization of innovation at the intersection of network control and information theory.

State dependent resource allocation is critical in wireless communications. For example, interference management (MIMO, interference alignment, etc) and opportunistic scheduling need accurate network state information from control channels. Remarkable progress has been made on the design of wireless networks with maximum throughput and, to some extent, for low latency by exploiting the opportunistic gain that comes from network and channel knowledge. However, most of these works assume that the control channel of wireless networks is perfect and that the network is fully observable. This assumption is becoming increasingly questionable because of two reasons: (i) the requirement of ultra-low latency make is it difficult (if not impossible) to get perfect (or exact) network state; and (ii) the multi-carrier technology and the ever-increasing size of wireless networks make it extremely expensive to obtain the complete (or full) network state information as the amount of information increases super-linearly while the capacity scales linearly.



It seems unavoidable that low latency wireless networks have to be controlled with incomplete and imperfect network state, which has a profound impact on the network performance. For example, consider an uplink network with two mobiles and a single base-station. Figure 1 depicts the throughput regions of three cases[1]: (i) complete and perfect information, where the base station has the instantaneous and perfect channel state information, (ii) delayed information, where the base-station has one time slot delayed channel state information; and (iii) information mismatch, where the base station has one time slot delayed information, but views it as the instantaneous information. We can see that the delayed channel state

[1] Lei Ying and S. Shakkottai. On Throughput Optimality with Delayed Network-State Information. In IEEE Transactions on Information Theory, Vol. 57, No. 8, Aug. 2011.

information leads to a throughput degradation. Furthermore, the information mismatch leads to a significant throughput loss. Therefore, it is important to understand the fundamental role of control channel on network performance. While the impact of limited feedback information on fundamental information theoretical capacity and network throughput has been studied, the impact on low latency transmissions is largely unexploited. To support low-latency communications in future wireless networks, we need to develop new and fundamental theories of delayed and incomplete network state information on low-latency communications, and equally importantly, we need to develop distributed and low complexity control algorithms based on imperfect and local network state information (with control channels with limited capacity) to achieve low latency communications.

Information theoretic tools have not been put to use to understand fundamental limits on control signaling that is required for efficient data communications. While such an understanding is not critical when latency is not critical, efficient design of control signaling becomes critical to enable efficient low latency communication schemes. For example, uncoordinated contention based access, while minimal in terms of control signaling, does not result in guaranteed low latency unless throughput efficiency is sacrificed significantly. Hence, more sophisticated medium access control strategies that involve some sophisticated control signaling are required. A fundamental understanding of what the minimum required amount of control signaling for different performance objectives and, in particular, to achieve low latency communication is required. We have been extending information theory to network control for decades but a complete theory has not yet emerged, especially in the context of low latency. This is a direction we need to work on to get asymptotic and non-asymptotic results. Similarly many questions on multi-user information theory are unanswered and these have implications on designing network control for low latency networks. In particular, we do not yet have a rigorous and fundamental understanding about the distributed communication model where users do not jointly optimize their channel codes. The distributed communication model may be important for low latency.

3.4 Information theory and Control of dynamical systems over wireless networks

In the last two decades, many control applications have emerged where a dynamical system is being controller by a remotely located controller. Perhaps, the most important of these is control of industrial plants (chemical processing, manufacturing, etc.), where a controller (either a human or a machine) observes the readings from different sensors and decides the operating point. Currently, the connections between the sensors and the controllers and between the controller and the actuator are wired, but if low-latency high-reliability links can be established, then wireless connectivity becomes attractive. In other emerging applications such as control of mobile agents (such as UAVs, submarines, vehicles, etc.), actuation and control must take place over wireless links.

3.4.1 State of the art

A fundamental tension emerges when one tries to combine information theory with control theory. Information theory does not take delay into account; control theory does not take communication delay or communication noise into account. Two distinct approaches have emerged to address these differences.

Stabilization over communication channels: The typical model in this setup consists of a dynamical system (also called a plant), a co-located sensor that observes the state of the plant and communicates to a remotely located controller over a communication channel; in some models, a communication channel between the controller and the actuator is also assumed. Various assumptions on the channels have been considered, including rate constraints, channel noise, packet drops, packet delays, etc. Most results share a common feature: the system is stabilizable if the sum of the log of unstable eigenvalues of the plant is less than a measure of channel quality (rate, Shannon capacity, anytime capacity, and variations thereof).

Optimal control and estimation over communication channels: The models considered in this setup are similar to those considered for stabilization, but the objective is to minimize a cost function rather than simply keep the state close to pre-specified trajectory. Most of the attention has been restricted to optimal estimation in this setup; the hope is that some form of separation theorem will take care of the control aspect although there are a few papers that consider joint estimation and control. Two approaches have been used for estimation (or zero-delay communication): coding of individual sequences and coding of Markov sources. In the former, explicit coding schemes are proposed that minimize the regret; in the latter, structure of optimal coding and decoding schemes is identified and dynamic programs to search for the optimal coding schemes is also identified.

3.4.2 Opportunities and future directions

Most of the models for stabilization over communication channels have considered a single plant connected to a controller over point-to-point communication channels. An important future direction is *to generalize these models to multiple plants connected to multiple controllers over multi-terminal communication channels*. Such a multi-terminal setup raises important cross-disciplinary challenges. In particular, what is the appropriate notion of decentralized stability of a plant, what is the appropriate notion of capacity for multi-terminal stabilization of plants, what is impact of the multiple access protocol on multi-terminal stabilization, etc.

Another important future direction is to consider *application layer approaches that trade-off latency with* freshness *of information*. In particular, for delay-sensitive applications such as control of dynamical systems and monitoring of time-series, one has to account for the network congestion caused by the data generated by an application. One possible approach is to keep track of *'update age'* to determine the rate at which new packets are generated; another approach is to keep track of *'information events'* to determine the time-instances when new packets are generated. There has been some preliminary analysis of these approaches, but a more detailed understanding of these trade-offs will provide a clearer picture of the role of communication network, and in particular of communication delay, for control of dynamical systems.

4. Ultra-Low Latency Wireless Applications

Today's communication networks and the Internet are largely geared towards moving latency tolerant (web, chat, email) and medium-latency (like voice) content. This has been driven by the fact that networks to date have been largely focused on personal communications and human perception and reflex reactions level out at close to 100 ms. Indeed, to handle increasing traffic load, existing wireless networks have been designed and planned with capacity and coverage in mind. The latency implications of the different applications have mostly been an after-thought.

However, new kinds of socially useful applications that use automated sensors and actuators working in closed-loop control systems are emerging in a range of domains. 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 much more demanding than the traditional applications involving humans at both ends, requiring two or three orders of magnitude improvement. Furthermore, these emerging applications often require deployment at a larger scale, in terms of both the numbers of nodes involved in (broadcast or multicast or convergecast) communications and area of operation, making it even more challenging and expensive to provide ultra-low latency over the network.

With respect to these various emerging low-latency applications, while for some applications it may be that general purpose existing (e.g., Cellular or WiFi) networks can be improved to support them, fundamentally new kinds of application-specific or customized networks may need to be designed to provide much tighter latency for which deeper technical research is needed on many fronts. Beyond purely technical challenges, moreover, there are also many regulatory and market incentive issues that need to be addressed to allow socially useful and innovative ultra-low latency applications to emerge in a fair and equitable manner.

4.1 Key Application Domains for Ultra Low Latency Wireless Networks

There are several domains of socially useful applications that would benefit greatly from enabling the operation of ultra-latency wireless networks:

4.1.1 ***Sensing and Actuation, Machine-to-Machine communications****:* There are many existing and emerging applications involving wireless networks of embedded sensors and actuators particularly in the context of industrial control in which latency is critical. Current standards have started to address these issues, for instance with the use of time-slotted, frequency scheduled protocols (e.g., WirelessHART, 802.15.4e TSCH/6tisch), however present wireless technologies typically provide only latencies on the order of milliseconds, which can be limiting for real-time control applications.

4.1.2 ***Human to Machine Communications****:* Remote equipment control is useful for applications that need remote motion/movement control of a machine over wireless network. Since such applications require a closed loop control, the latency requirements of the wireless part are quite stringent especially when the person performing the control is geographically distant from the machine. Example applications include controlling a robotic arm over the Internet.

4.1.3 ***Vehicle-to-Vehicle Communications***: This is a rapidly emerging domain of applications where there is greater mobility, dynamics, fluctuation in network size and topology. Application areas that require low latency are vehicular safety applications and traffic efficiency applications. Today's systems utilizing 802.11p or LTE-direct standards are not primarily designed for ultra-low-latency. Applications like automated lane changing, crash warning and autonomous-response systems require guaranteed ultra-low latency, which remain a challenge.

4.1.4 ***Parallel & distributed computing, massive MIMO, mobile off-loading**:* These applications are encountered in, for example, data centers with real-time map-reduce-type computation-heavy workloads, and for the backhaul of raw signals from base stations to compute centers for cooperative communications (distributed multiuser / massive MIMO) for next generation cellular systems. Further, ultra-low-latency off-loading of computation from mobiles to a data center could allow computationally much richer applications to display seamlessly on mobile devices.

4.1.5 ***Immersive services***: Online gaming, augmented reality, 3D/holographic virtual reality, brain-to-brain communication are some examples of next-generation immersive applications involving potentially large collections of humans communicating and collaborating over large geographical distances that can greatly benefit from ultra-low latency networking. Even today, online multiplayer mobile gaming over cellular network is far from a reality in spite of the advances of 4G LTE technology.

4.1.6 ***Wireless networks on chip for health monitoring and control:*** Portable, miniaturized and even tissue friendly sensors for monitoring biological processes within and at the skin surface become day by day not only available but also capable of sensing and producing significant amount of biological data. For instance, it is possible to sense and measure blood glucose and metabolic products from skin and soft tissue secretions offering non-invasive information about human physiology. Yet, integrative computational systems capable of mining biological processes from measured bio-chemical reactions to proteins and physiological processes and determining more accurate therapeutic strategies are missing. Ultra-low latency wireless communication within a chip and between multiple chips can enable not only the transmission of sensed data but can also enable the control at molecular level by triggering molecular and cellular actuators. In addition, wireless based NoCs can support the mining and analysis of biological data identifying patterns of abnormality and determine control strategies.

4.2 Classification of Ultra-Low Latency Applications

In addition to obvious latency constraints, ultra-low latency applications have additional characteristics that impose challenges in network design. These characteristics could be viewed as additional dimensions in the design space along with latency dimension. To better understand the additional requirement of low-latency applications, we classify the low-latency applications along many dimensions in the following.

4.2.1 Machine (M) versus Human (H) endpoints

There are 4 distinct possibilities: H2H, H2M, M2H, and M2M. Applications that involve humans such as online gaming, are relatively delay tolerant in that humans cannot resolve delays of less than 10ms. For example, individual video frames at 30 frames/sec (or 30 ms between frames) cannot be resolved by the human vision system. M2H and H2M applications have similar requirements because of the human interaction. By contrast, M2M systems have delay requirements that may be arbitrarily tight. For example, vehicle-to-vehicle (V2V) safety messaging may require 1 ms delays as a 57 mph car moves at 1 inch/ms and inches matter in controlling the position of cars in a heavy traffic. As new M2M applications emerge, latency requirements may become even more stringent.

4.2.2 Scale of application (number of nodes, spatial extent, topology, heterogeneity)

Low-latency networked systems may range from a single short-range communication link (one sensor sending to a monitor) to systems of arbitrarily large size and geographic scope. Vehicle-to-vehicle systems may include several hundred cars on a highway. Massive multiplayer online gaming may involve thousands of players spread around the world communicating through wireless and wired network links. These examples also highlight that topologies can be local, metropolitan, or global.

4.2.3 Open loop versus closed loop

Many M2M industrial control applications will need closed loop systems with timely packet acknowledgements. In other applications, notably V2V safety messaging, much of the signaling will be open-loop. Cars will broadcast messages but not every car in the vicinity will receive each message. However, low latency may still be important. In such open loop systems, system design will need to accommodate the absence of feedback. This is related to reliability requirements.

4.2.4 Latency and Reliability requirements

In many applications, low latency is merely desirable. For example, in H2H immersive services, violating latency requirements results in a degraded quality of experience, which is undesirable but tolerable. In other environments, latency failure can cause substantial damage. Actuators controlling an electric power plant can induce systems failures if sensor reports are not received in time. Similarly, unreliable delivery of messages in a remote operated vehicular control system may cause people to be injured.

4.2.5 Tradeoffs with other metrics: Throughput, Age, Energy, Reliability, Spectral Efficiency, Security, Cost, Scale (nodes, spatial extent, topology)

In the context of specific systems, it is generally accepted that more stringent latency constraints will penalize other performance metrics; throughout and spectral efficiency will be reduced while spatial extent and system scale are likely to be restricted. These effects would cause system costs to rise. The precise nature of these tradeoffs will depend on the specific applications. (With respect to theory, it is not clear that these recognized tradeoffs are fundamental. Performance bounds on latency that apply across all systems are not known.)

4.3 Socio-economic issues concerning ultra-low latency wireless applications and their adoption

To generate real value, ultra low-latency applications must be accessible by end-consumers. However, these applications currently face a “catch 22.” On the one hand poor network performance in many areas of the country remains a barrier to innovation, while on the other hand without a healthy market for ultra-low latency applications, there is not enough of an incentive to develop and deploy networks that support ultra low-latency communication.

Driven by industry interests a few applications were able to break this cycle and create incentives for improvements in network latency. To support low latency needs of Web applications, research on Content Distribution Networks (CDNs), front-end servers, and datacenter networks has improved end-to-end user request delay. Similarly, high speed trading has incentivized ISPs to establish low-latency paths between several cities. However, many emerging application may neither have the large customer base of Web applications, nor the financial influence of electronic traders to motivate network operators to drive latency down. While these large applications do stimulate investment in Internet infrastructure and do drive latency down in certain areas, these gains do not go far enough to meet the needs, both in terms of latency and price, of new types of application waiting in the wings.

4.3.1 Effect of pricing mechanisms on provisioning of low latency network services

Most ISPs advertise and sell services based on network bandwidth, not latency. Although there are exceptions to this rule for enterprise customers, for example high-speed connections via MPLS circuits, residential and mobile broadband services do not have guarantees on end-to-end, or last-mile, latency. As a result ISPs tend to trade off latency to increase throughput in their networks. Examples of such practices include delay of transmission in cellular schedulers until high RSSI supports efficient spectrum utilization, bufferbloat at routers to support high link utilization, or use of circuitous Internet forwarding paths to leverage inexpensive peering agreements. One could argue that continual increase in network capacity will eliminate network congestion and network delay with it. However, dramatic improvements to network capacity from advances in physical and link layer technologies have not had a commensurate impact on network latency.

There are two major reasons for this stagnation. First, network capacity is deployed in response to growing traffic demands, which tend to fill up a growing network. The deployment of new technologies, notably nationwide 4G networks, is followed by advertising campaigns, which bring new network customers, whose traffic helps fill out the new capacity and provide return on investment. As a result the effect of new capacity on queue reduction and delay are often short lived. The second reason for the slow pace of latency reduction are long packet forwarding paths, which result from increasingly convoluted network configuration and management practices. For example cellular networks forward packets through a series of aggregation nodes and accounting middleboxes that inflate packet delay to tens and even hundreds of milliseconds even before packets reach the gateway router. Further, circuitous forwarding paths used by some rural ISPs can bounce traffic between several cities without monotonic progress towards a destination.

4.3.2 Implications of Network Neutrality regulation

The latest Net Neutrality Order announced by the Federal Communications Commission (FCC) establishes regulation of broadband providers, including cellular network operators, under the Title II of the Communications Act. Though the exact regulation is still being decided, this classification allows the FCC to enforce three guiding principles on customer traffic: no blocking, no throttling, and no paid prioritization. These rules are intended to prevent anticompetitive behavior and what public perceived as extortion by some network operators. Although this decision to regulate broadband under Title II has gained much praise for open Internet advocates, there may be some potential implications for the adoption of low latency communications stemming from that decision.

The benefits of prioritization for some types of traffic, for example voice, have long been seen as beneficial both due to the need for low end-to-end latency of voice communication and the relatively low volume of individual flows. As such VoIP services are exempt from the no paid prioritization clause as "specialized services." Although FCC allows network operators to provision specialized services differently from broadband services (even when they carry IP traffic), specialized services may not carry traffic that can otherwise be perceived as broadband. As such, some ISPs have refused customer requests to carry streaming video traffic configured as a specialized service. The rigidity of specialized service classification leaves many other types of latency sensitive traffic in the common pipe. Although in many ways that represents the status quo, it also perpetuates many of the challenges of low latency traffic in networks configured for bandwidth maximization (described in the preceding section). At the same time the FCC seems to leave a door open for prioritization of traffic, as long as it is not paid prioritization. As part of the recent ruling the FCC has also set up a mechanism for network operators to seek opinions on changes to their network configuration practices.

What could such FCC-approved prioritization, or preferential treatment of latency-sensitive traffic, look like? In the interest to improve user experience and to support emerging classes of low-latency applications network operators may want to support low latency network services, again as long as such prioritization is not paid. Network services that provide low latency, but also restrict bandwidth could be a viable model that does not result in delay tolerant, high volume traffic clobbering the narrower channel. Other approaches could include application-specific network configurations, which however may be harder to manage from the network perspective. Recent advances in the SDN and network virtualization technologies, however, may make such management more practical than older Intserv and Diffserv approaches. At the moment it is not clear whether network operators would have a business case to provide any such prioritization subject to FCC approval, however the alternative of a common pipe may prove challenging for the emergence of new low-latency applications. Network economics analyses may be needed to shed a useful light on optimal regulatory policies to encourage innovation in low-latency applications.

4.3.3 Equity of access to low latency network services

For ultra low-latency applications to reach a broad audience, low latency needs to be achievable in all geographic areas and at a low cost. Network operators tend to invest in new technologies in urban areas, where they are likely to see a quick return on investment and larger profits. Under the business model that tries to maximize network capacity such investment makes sense, because urban networks are more likely to suffer from congestion induced delays. Rural networks on the other hand tend to suffer from long Internet paths and poor signal strengths due to sparse deployment of cellular infrastructure. Network services might need to be regulated in subsidized to prevent a new type of digital divide, in which low latency applications are not available in rural areas. On the other hand there is hope for industry forces to demand broad, low latency coverage. For example remote operation of drilling, mining, and farming equipment requires low latency interactive communication between equipment and remote operators. Because such equipment is likely to operate with side-by-side humans, real-time control is critical to its safe operation. Although industrial applications of low latency communications may drive network performance improvement in some areas, it may need additional regulation to make sure these advances reach general population in a timely and permanent manner.

Appendix A: Workshop Participants

Randall Berry , Northwestern University
Paul Bogdan , USC
Koushik Chakraborty, Utah State University
Supratim Deb, Alcatel-Lucent
Anthony Ephremides, University of Maryland
Atilla Eryilmaz, Ohio State University
Amlan Ganguly, RIT
Deuk Hyoun Heo, Washington State University
I-Hong Hou, Texas A & M University
Jing Jiang, Qualcomm
Ashish Khisti, U. Toronto
Avinash Kodi, Ohio University
Bhaskar Krishnamachari, USC
Jie Luo (Rockey), Colorado State University
Aditya Mahajan, McGill
Vivek Mahtre, AT&T
Radu Marculescu, CMU
Eytan Modiano, MIT
Borivoje Nikolic, UC Berkeley
Steve Nowick, Columbia University
Umit Ogras, Arizona State University
Partha Pande, WSU
Craig Patridge, Raytheon BBN
Ram Ramanathan , Raytheon BBN
Ashu Saberwal, Rice University
Anant Sahai, UC Berkeley
Saikat Sarkar, Broadcom
Sanjay Shakkottai, University of Texas Austin
Baris Taskin, Drexel University
Leandros Tassiulus, Yale
Manos Tentzeris, Georgia Tech
Harish Viswanathan, Alcatel-Lucent
Mike Wittie , Montana State University
Ian Wong , National Instruments
Roy Yates, Rutgers University
Edmund Yeh, Northeastern University
Lei Ying, Arizona State University
Charlie Zhang, Samsung
Junshan Zhang, Arizona State University
Ting Zhu, UMBC

Appendix B: Workshop Agenda

Thursday, March 26:

8:00 Breakfast

8:30 opening remarks

8:45 – 10:00 Short presentations

Radu Marculescu (CMU)
Harish Viswanathan (Alcatel-Lucent)
Ashu Sabrawal (Rice)
Sanjay Shakkottai (UT-Austin)
Randall Berry (Northwestern University)

10:00 – 10:30 Break

10:30 – 12:30 Breakout sessions

Network control
Wireless Systems
On-chip
Information theory

12:30 – 1:30 Lunch

Talks by Roy Yates and Saikat Sarkar

1:30 – 3:00 Report back from breakout sessions

3:00 - 3:30 break

3:30 – 5:30 Breakout sessions with mixed participants

7:00 Dinner

Friday, March 27

8:00 Breakfast

8:30 – 9:30 Report back from afternoon breakout sessions

8:30 - 9:30 Small group convenes to plan out the report

9:30 - 10:00 Break

10:00 – 11:30 Breakout session

11:30 Report plan and writing assignment

12:00 Lunch

12:30 – 3:00 Breakout session for report writing