Throughput and Latency of mmWave Communications

Assoc. Prof. Ming Xiao, Jointly with Dr. Guang Yang

Department of Information Science and Engineering (ISE)
KTH Royal Institute of Technology, Stockholm, Sweden

September 14, 2018

Acknowledgment:

Wireless@KTH Seed Project: Millimeter-Wave for Ultra-Reliable Low-Latency Communication
Introduction
5G Communications

Figure: IMT-2020 vs. IMT-Advanced

8 KPI in 5G
- user experienced data rate
- peak data rate
- area traffic capacity
- network energy efficiency
- connectivity density
- latency
- mobility
- spectrum efficiency

---

3 Categories of Use Scenarios

- **eMBB**: unprecedented data volumes, overall data capacity and user density
- **mMTC**: low power consumption & data rates for massive connected devices
- **URLLC**: safety-critical and mission-critical applications

**Figure**: 3 categories of application scenarios

---

2[https://www.etsi.org/technologies-clusters/technologies/5g](https://www.etsi.org/technologies-clusters/technologies/5g)
Introduction

mmWave in 5G

5 Important Enablers of 5G Systems

- scalable OFDM-based air interface
- flexible slot-based framework
- advanced channel coding
- massive MIMO
- mobile mmWave

Figure: use cases of mmWave

---

Introduction
mmWave in 5G

Features of mmWave

- **abundant bandwidth resources** (spectra from 30 GHz to 300 GHz)
  - enable high area & per-user data rates
  - enable flexible spectral resource allocations
- **very short wavelength** (wavelength from 10 mm to 1 mm)
  - allow dense antenna array for sharp beams
- **high penetration loss & atmospheric absorption**:
  - recommended for short-range scenarios
- **weak diffraction & reflection**
  - interference-free spatial reuse
Various Definitions for Latency & Reliability

Latency

- **end-to-end latency**
  including propagation delay, queuing delay, processing/computing delay and retransmissions

- **user-plane latency (3GPP)**
  one-way time it takes to successfully deliver a packet

- **control-plane latency (3GPP)**
  transition time from idle state to active state

Reliability

- **reliability (3GPP)**
  probability of finishing transmission within given time duration

- **reliability per node**
  probabilities of transmission error, queuing delay violation probability, or proactive packet dropping

- **control channel reliability**
  probability of successfully decoding the scheduling grant or metadata

- **availability**
  probability of available service

Introduction

URLLC with mmWave

Figure: key enablers for low latency and high reliability

- Short TTI
- Caching
- Densification
- Grant-free
- UAV/UAS
- Non orthogonal multiple access (NOMA)
- MEC/FOG/MIST
- Network coding
- Machine learning
- Slicing

Reliability \((1 - 10^{-\sigma})\)

Latency (\(ms\))

Best Effort

ULRCC

Low-Latency Communication (LLC)

Ultra-Reliable Communication (URC)

ENABLERS

- Short TTI
- Spatial diversity
- Network coding
- Caching, MEC
- Multi-connectivity
- Grant-free + NOMA
- Machine learning
- Slicing

ENABLERS

- Finite blocklength
- Packet duplication
- HARQ
- Multi-connectivity
- Network coding
- Spatial diversity
- Slicing

 ITS

Factory 2.0

September 14, 2018 2 / 19
Enablers for URLLC
- short transmission time interval
- spatial diversity
- high data rates
- network coding
- caching/mobile edge computing
- multi-connectivity
- grant-free NOMA
- machine learning
- slicing

Benefits from mmWave
- spatial diversity
  - sharp beams
  - interference-free spatial reuse
- high data rates
  - large bandwidth
  - less interference
- multi-connectivity
  - flexible spectrum allocations
  - sharp beams
  - interference-free spatial reuse
- grant-free
  - abundant bandwidth resources
Our related contributions:

1. **throughput of relay network with directional antennas**
   - *Performance Analysis of Millimeter-Wave Relaying: Impacts of Beamwidth and Self-Interference* [TCOM’18, published]

2. **latency of buffered networks with different architectures**
   - *Low-Latency Millimeter-Wave Communications: Traffic Dispersion or Network Densification?* [TCOM’18, published]

3. **traffic allocation for low-latency buffered networks**
   - *Traffic Allocation for Low-Latency Multi-Hop Millimeter-Wave Networks with Buffers* [TCOM’18, published]

above as parts of outcomes from **Wireless@KTH Seed Project**

*Millimeter-Wave for Ultra-Reliable Low-Latency Communication*

for more information, please see the project homepage:

https://sites.google.com/view/wirelesskth-urll-mmwave/home
Performance Analysis of Millimeter-Wave Relaying: 
Impacts of Beamwidth and Self-Interference
Introduction

Background

- directional antenna is enabled by short wavelength to combat path loss
- relay can be adopted to improve communication range

Related Work

- sectorized model for directional antenna
  - highly tractable for theoretical analysis
  - miss out “roll-off” feature in realistic antennas
- no reflection incorporated in mm-wave channel model
  - greatly simplified mm-wave link models
  - first-order reflection indeed matters (shown by measurements)
Introduction

Motivation: to address limitations of models in preceding works
- Two-Ray Channel Model: take reflections into account
- Gaussian Antenna Model: capture the practical “roll-off” feature

Contribution
- investigate performance of AF relays in two duplex modes
- characterize impacts of beamwidth and self-interference
- show non-trivial contribution of ground reflections
System Model
Two-Hop AF Relaying System

Figure: Two-hop AF relaying system: two-ray channel and directional antennas.

System Setting
- linear network deployment and no signal from $S$ to $D$
- Gaussian directional antenna
- half-duplex (HD) or full-duplex (FD) AF relay, and self-interference for FD
- sum-power constraint to $S$ and $R$
given sufficiently high sum power, direct transmission is the best option

self-interference significantly affects performance of FD relaying

FD relaying with small self-interference should be adopted for medium sum-power region

HD relaying performs better for low sum power

Figure: Rates vs. sum-power constraint, where $L_1 = 80$ m and $\theta_m = \frac{\pi}{6}$.

Ming Xiao (KTH)
1-ray and 2-ray models may differ a lot when beam is wide.

2-ray model reduces to 1-ray model, approximately, when using sharp beams.

Reflections constructively or destructively contribute to performance, depending on location of relay deployment.

Figure: Contribution of ground reflections, where $\xi = 100$ dB.
Performance Evaluation

- performance of FD relaying convexly degrades with ascending self-interference
- given higher sum power, benefit by reducing self-interference grows

Figure: Impact of $\mu$ on $\eta_{FD}^*$, where $\theta_m = \frac{\pi}{4}$ and $L_1 = 100$ m.
Concluding Remarks

- FD relaying outperforms its HD counterpart, only when sharp beams and smaller self-interference coefficient are applied.

- Given sufficiently high sum-power budget or sharp beams, direct transmission outperforms two relaying schemes, then not necessary to use AF relaying.

- Impacts of beamwidth and self-interference coefficient can be characterized by $O\left(\min\{\theta_m^{-1}, \theta_m^{-2}\}\right)$ and $O\left(\mu^{-\frac{1}{2}}\right)$, respectively.

- Ground reflection may significantly affect the performance of mm-wave communications, constructively or destructively.
Low-Latency Millimeter-Wave Communications: Traffic Dispersion or Network Densification?
Introduction

Background
- low latency is critical QoS feature for delay-sensitive applications in 5G
- buffers are adopted to support heavy network traffic
- queuing delay largely affects total delay in buffered systems

Related Work
- numerous efforts addressing low latency
- however, inadequate works from following two aspects:
  - buffer-aided scenarios are rarely discussed in preceding works
  - limited research from perspective of network architectures
Introduction

Motivation: study latency in buffered systems in terms of architectures

- traffic dispersion and network densification are effective schemes

- potentials of traffic dispersion and network densification are not clear yet

- combination of above schemes may be promising for certain scenarios

Contribution

- investigate the respective strengths of traffic dispersion and network densification, and propose a generic hybrid scheme

- closed-form MGF-based service characterization for mm-wave system with Nakagami-$m$ fading is obtained in $(\min, +)$ algebra (instead of $(\min, \times)$)

- bounds for probabilistic delay and effective capacity are derived for performance evaluation
System Model
Traffic Dispersion

**Mechanism:** reduce arrival rate (per path)
- partition original arrival traffic into multiple sub-streams
- each sub-stream is served and delivered to own path, independently
- receiver collects all incoming sub-streams to form final output traffic

**Strength**
- offload arrival traffic to avoid large queue in single buffer

**Figure:** Illustrations of traffic dispersion

![Traffic Dispersion Diagram](image-url)
**Mechanism:** increase service capability
- increase density of intermediate nodes to shorten hop length
- output traffic from one node acts as input traffic for next node

**Strength**
- increase per-hop service capability to improve end-to-end performance
System Model

Hybrid Scheme

**Mechanism**: offload arrival traffic + increase per-hop service capability
- flexible combination of *traffic dispersion* and *network densification*

**Strength**
- combine respective strengths of traffic dispersion and network densification

*Figure*: Illustration of hybrid scheme
Latency-Related Metrics

Violation Probability

\[ p_w \triangleq \Pr (W(t) \geq w) \]

- \( W(t) \): stochastic latency
- \( w \): given latency threshold
- \( p_w \): violation probability, lower is desirable

Effective Capacity

\[ C(-\theta) \triangleq -\lim_{t \to \infty} \frac{\log(\mathbb{E} \left[ \exp(-\theta \sum_{i=1}^{t} C_i) \right])}{\theta t}, \text{ for } \theta > 0 \]

- \( C_i \): instantaneous channel capacity
- \( \theta \): QoS exponent, indicating stringent requirement for high \( \theta \)
- \( C(-\theta) \): effective capacity, higher is desirable
bounds accurately capture trends of violation probability for both schemes
more parallel channels or higher relay density is always beneficial.

Figure: $\epsilon^w$ vs. $w$ with $\rho = 2$ Gbps and $\gamma = 85$ dB.
Traffic dispersion performs better in higher sum-power region.

Network densification performs better in lower sum-power region.

Lower & upper bounds for effective capacity of network densification are very close.

**Figure:** $C(-\theta)$ vs. $\gamma$ with $\theta = 2$ (U.B./L.B. denote upper/lower bound)
Figure: $C(-\theta)$ vs. $m$ with $n = 12$ and $\theta = 2$

- role of traffic dispersion becomes more important when sum-power budget grows
- proper number of paths & per-path relays should be adopted for medium sum-power scenarios
Concluding Remarks

- Traffic dispersion, network densification, and hybrid scheme show own strengths, with respect to high, lower and medium sum powers, respectively.

- It is always beneficial to have more independent paths or higher relay density, while resulting gain heavily depends on sum-power budget.

- Bounds derived for probabilistic delay and effective capacity work well for performance assessment.
Traffic Allocation for Low-Latency Multi-Hop Networks with Buffers
Introduction

Background

- low latency is critical feature in 5G mobile communications
- proper traffic control can mitigate network latency

Related Work

- latency of buffered multi-hop network is extensively investigated
- however, limited works dedicated towards studying:
  - traffic allocation for low latency rarely discussed previously
  - buffers not incorporated when studying traffic allocation
Introduction

Motivation

- unavailable works analyzing buffered systems via graph-based approaches
- multi-hop buffered network with multiple parallel channels per hop is not discussed in preceding works

Contribution

- investigate two traffic allocation schemes, i.e., local/global allocation
- exploit recursive nature of global allocation, and show the overall computational complexity for local and global schemes.
- perform asymptotic analysis and derive lower bounds for latency when considering Nakagami-$m$ fading mm-wave channels
### System Model

**Figure:** A multi-hop network with multiple channels in each hop.

### System Setting

- each FD relay is equipped with one buffer
- one fraction is partitioned and forwarded, unless completely received
- any fraction is infinitely divisible
System Model

Procedure of Traffic Allocation

(i) incoming traffic is first partitioned into several smaller fractions according to the given allocation scheme

(ii) fractions are subsequently pushed onto channels and delivered to next node, where each fraction can be partitioned again for further delivery.

Traffic Allocation Schemes

\(M_{\text{local}}\): node \(h\) only has capacity information of channels in hop \(h-1\), and only transmission over hop \(h-1\) is optimized

\(M_{\text{global}}\): node \(h\) has capacity information of all channels from hop 0 to \(h-1\), and transmission from hop 0 to \(h-1\) is optimized

Definition (End-to-End Latency \(\tau_n\))

time span for delivering one fixed-length file of size 1, from the moment the source starts transmission to the moment all fractions are received at destination
Performance Evaluation

Figure: Minimum latency $\tau^*$ vs. number of relays $n$ (with number of channels $m$)

- more relays gives higher latency
- increasing number of channels can significantly reduce latency of two schemes
- advantage of global scheme is growingly significant as number of relays increases
merits of having multiple channels vanish as number of channels increases

Figure: Minimum latency $\tau^*$ vs. number of channels $m$ (with number of relays $n$)
Concluding Remarks

- Comparing $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$, lower latency is achieved albeit higher computational complexity.

- More parallel channels give lower latency, while benefit diminishes as number of channels increases.

- Proper deployment of relay nodes, which may be different for distinct schemes, is critical role in minimizing latency.