

[x-area] AI/ML Traffic

eMBB consumer

MIMO

- CSI enh.
- BM: [subject to R17]
- Stationary: 8Rx, overhead redux
- UL sub-band precod.
- UL 4+ layers

DC/CA Enh.

- X-carrier HARQ: feedback & re-Tx
- Fast re-Tx split bearer
- Temporal RS PScell act
- Scalable x-carrier sch.

XR/CG Enh.

- QoS+, x-layer opt.

MBS

- SFN+
- QoS+ (Tput, reliab.)
- TV (ATSC3.0 ref)

NW Topology

Sidelink LLeMBB

- SL-U esp. <7GHz, FR2
- Low latency 1Gbps
- SL-U RedCap

Sidelink Relay

- U2U relay
- UE scheduling UE
- mPath, mHop
- Mobility (Remote, Relay)
- Network coding

Smart Repeaters

- Beamforming
- Interf. Mgmt (T/F DD)
- Integration (UE authorization)

NTN Evolution

NTN NR

- Mobility
- Regenerative arch
- HD-FDD, VoNR, MBS
- R17 leftovers

NTN IoT

- Mobility (connected)
- R17 leftovers

SID Spectr. sharing

- Study scenarios, target spectrum and regulation status

Long-term explor.

SID AI/ML integr.

- NG-RAN/AS integrat.
- DMRS ch. est., Rx noise suppress, CSI-RS overhead, CSI feedback
- (UE-based) Mobility predict., Pos. enh.
- NW functions (load balancing, radio resource planning..)

SID AI traffic

- Traffic and arch.
- Overhead optim.

SID >71GHz

- Spectrum charac.

Common tech.

[FR2] Mobility

- L1/L2 trig. CHO
- Inter-/intra-cell beam switching delay redux
- RRC DAPS HO mPanel

System Energy

- DCI-based pwr sav mTRP and mPanel
- gNB/TRP dormancy (UE -trig. / -imposed)
- Eval. Methodology (Pwr. Cons. Models)

POS (NR, SL, RedCap)

- cm-level (Tx + meas related to signal ϕ)
- SL (-based, -assisted)
- RedCap UE
- R17 leftovers

SID gNB Full Duplex

- Partitioning, scenarios, interf.

Verticals

URLLC

- DL control efficiency
- NR-U enh

RedCap

- PA-less
- (POS)
- NO LPWA

(UAV: neutral)

eMBB

MIMO

- CSI enh.
- BM: [subject to R17]
- Stationary: 8Rx, overhead redux
- UL sub-band precod.
- UL 4+ layers

DC/CA Enh.

- X-carrier HARQ: feedback & re-Tx
- Fast re-Tx split bearer
- Temporal RS PScell act
- Scalable x-carrier sch.

Sidelink LLeMBB

- SL-U esp. <7GHz, FR2
- Low latency 1Gbps
- SL-U RedCap

XR/CG Enh. [SA-led]

- QoS+, x-layer opt.

NTN NR

- R17 leftovers
- Mobility
- Regenerative arch
- VoNR, MBS, HD-FDD

MBS

- SFN+
- QoS+ (Tput, reliab.)
- TV (ATSC3.0 ref)

(may also be seen as non-eMBB)

Non-eMBB

URLLC

- DL control efficiency
- NR-U enh

RedCap

- PA-less
- (POS)
- NO LPWA

NTN IoT

- R17 leftovers
- Mobility (connected)

(UAV: neutral)

X-areas New areas

System Energy

- DCI-based pwr sav mTRP and mPanel
- gNB/TRP dormancy (UE -trig. / -imposed)
- Eval. Methodology (Pwr. Cons. Models)

[FR2] Mobility

- L1/L2 trig. CHO
- Inter-/intra-cell beam switching delay redux
- RRC DAPS HO mPanel

Sidelink Relay

- U2U relay
- UE scheduling UE
- mPath, mHop
- Mobility (Remote, Relay)
- Network coding

Smart Repeaters

- Beamforming
- Interf. Mgmt (T/F DD)
- Integration (UE authorization)

POS (NR, SL, RedCap)

- cm-level (Tx + meas related to signal ϕ)
- SL (-based, -assisted)
- RedCap UE
- R17 leftovers

SID NTN f sharing

- Study scenarios, target spectrum and regulation status

SID gNB Full Duplex

- Partitioning, scenarios, interf.

SID AI/ML integr.

- NG-RAN/AS integrat.
- DMRS ch. est., Rx noise suppress, CSI-RS overhead, CSI feedback
- (UE-based) Mobility predict., Pos. enh.
- NW functions (load balancing, radio resource planning..)

SID AI traffic

- Traffic and arch.
- Overhead optim.

Study AI/ML Traffic

RAN2-led

To characterize AI traffic and investigate AI traffic management
Efficient encoding of AI model and messages (“AI-Encoding”)

Objective I: AI-Encoding for non-split learning methods [RAN2, (3)]

- Study learning method and analyze data redundancy associated with AI/ML models
- AI traffic compression, e.g. AI model encoding, sparsification and quantization

Objective II: AI-Encoding for split learning methods [RAN2, (3)]

- Model splitting between UE and Server, especially for federated learning
- Model coding for combination of different model compression schemes, e.g. lossy/lossless, quantization, sparsification, and encoding.
- Compression in both training and inference

3GPP TUs (Total w/ 9 meetings)			
RAN1	RAN2	RAN3	RAN4
-	8	TBD	-

SA/CT Dependency: **Yes**

AI/ML Traffic Use cases

Observation 1: AI adoption has significantly accelerated during the pandemic across all major industries esp. Tech, Finance and Retail (source: KPMG [link](#))

Observation 2: AI/ML traffic will be a significant portion of mobile network traffic

It is important to ensure that the 5G System (Core and Access) can fully accommodate such traffic

AI/ML Traffic Study

SA1

- [SP-191040](#) Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS (FS_AMMT) [22.874](#)
- SA1 Study categorized use cases into three aspects:
 - AI/ML operation splitting between AI/ML endpoints
 - AI/ML model/data distribution and sharing over 5G system
 - Distributed/Federated Learning over 5G system.
- **Observation 3:** SA1 study identified following characteristics for AI/ML traffic
 - Training data can be distributed
 - Learning can be distributed
 - Model can be distributed and split
- **Proposal1:** Based on SA1 study, further study on AI traffic and required architecture for AI/ML traffic management.

AI/ML Traffic Study

- Communication overhead for AI/ML model or *intermediate data* transfer could be a show-stopper for including AI in Wireless
 - Efficient AI Traffic Management is needed
- AI-encoding for Efficient Communications
 - Study and analyze data overhead associated with AI/ML models
 - Role of model coding for reducing AI/ML model overhead
 - Role of model sparsification and quantization methods in traffic overhead reduction
 - Comparison and tradeoffs between lossy/loss-less encoding, sparsification and quantization methods
- Model coding and splitting between UE and network
 - Model coding and splitting for federated learning
 - Combination of different model encoding, sparsification and quantization schemes
 - Encoding in both training and inference parts of AI/ML for further overhead reduction

Proposal2: Study mechanism, e.g. encoding, to optimize AI/ML model transfer overhead.

Conclusions

Observation 1: AI adoption has significantly accelerated during the pandemic across all major industries esp. Tech, Finance and Retail (source: KPMG [link](#))

Observation 2: AI/ML traffic will be a significant portion of mobile network traffic

Observation 3: SA1 study identified following characteristics for AI/ML traffic

- Training data can be distributed
- Learning can be distributed
- Model can be distributed and split

Proposal 1: Based on SA1 study, further study on AI traffic and required architecture for AI/ML traffic management.

Proposal 2: Study mechanism, e.g. encoding, to optimize AI/ML model transfer overhead.

Thank You!

MediaTek TDocs to RAN Rel-18 Workshop

RWS-210092	MediaTek Views on Rel-18 content	MediaTek Inc.
RWS-210093	[eMBB] MIMO Enhancements	MediaTek Inc.
RWS-210094	[eMBB] DC/CA Enhancements	MediaTek Inc.
RWS-210095	[eMBB] XR/CG Enhancements	MediaTek Inc.
RWS-210096	[eMBB/Other] MBS Enhancements	MediaTek Inc.
RWS-210097	[eMBB] Sidelink Enhancements - LLeMBB	MediaTek Inc.
RWS-210100	[eMBB] NTN NR Enhancements	MediaTek Inc.
RWS-210101	[non-eMBB] NTN IoT Enhancements	MediaTek Inc.
RWS-210108	[non-eMBB] URLLC Enhancements	MediaTek Inc.
RWS-210109	[non-eMBB] NR RedCap Enhancements	MediaTek Inc.
RWS-210098	[x-area] Sidelink Relay Enhancements	MediaTek Inc.
RWS-210099	[x-area] Smart Repeaters Enhancements	MediaTek Inc.
RWS-210102	[x-area] NTN/TN Spectrum Sharing	MediaTek Inc.
RWS-210103	[x-area] AI/ML Integration	MediaTek Inc.
RWS-210104	[x-area] AI/ML Traffic	MediaTek Inc.
RWS-210105	[x-area] Mobility Enhancements	MediaTek Inc.
RWS-210106	[x-area] System Energy Enhancements	MediaTek Inc.
RWS-210107	[x-area] Positioning Enhancements	MediaTek Inc.
RWS-210197	[x-area] Sub-band Full-duplex for gNB	MediaTek Inc.
RWS-210110	Draft WID: System Energy Enhancements	MediaTek Inc.
RWS-210111	Draft WID: Mobility Enhancements	MediaTek Inc.
RWS-210112	Draft WID: DC/CA Enhancements	MediaTek Inc.
RWS-210113	Draft WID: NTN IoT Evolution	MediaTek Inc.