3GPP TR 22.876 V19.0.0 (2023-06)

Technical Report

3rd Generation Partnership Project;

Technical Specification Group Services and System Aspects;

Study on AI/ML Model Transfer Phase 2

(Release 19)



The present document has been developed within the 3rd Generation Partnership Project (3GPP TM) and may be further elaborated for the purposes of 3GPP.  
The present document has not been subject to any approval process by the 3GPPOrganizational Partners and shall not be implemented.  
This Report is provided for future development work within 3GPPonly. The Organizational Partners accept no liability for any use of this Specification.  
Specifications and Reports for implementation of the 3GPP TM system should be obtained via the 3GPP Organizational Partners' Publications Offices.

Keywords

AIML-Ph2

***3GPP***

Postal address

3GPP support office address

650 Route des Lucioles - Sophia Antipolis

Valbonne - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Internet

http://www.3gpp.org

***Copyright Notification***

No part may be reproduced except as authorized by written permission.  
The copyright and the foregoing restriction extend to reproduction in all media.

© 2023, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).

All rights reserved.

UMTS™ is a Trade Mark of ETSI registered for the benefit of its members

3GPP™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners  
LTE™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners

GSM® and the GSM logo are registered and owned by the GSM Association

Contents

Foreword 5

Introduction 5

1 Scope 6

2 References 6

3 Definitions, symbols and abbreviations 7

3.1 Definitions 7

3.2 Symbols 8

3.3 Abbreviations 8

4 Overview 8

5 Split AI/ML operation between AI/ML endpoints for AI inference by leveraging direct device connection 8

5.1 Proximity based work task offloading for AI/ML inference 8

5.1.1 Description 8

5.1.2 Pre-conditions 9

5.1.3 Service Flows 9

5.1.4 Post-conditions 11

5.1.5 Existing features partly or fully covering the use case functionality 11

5.1.6 Potential New Requirements needed to support the use case 11

5.1.6.1 Potential Functionality Requirements 11

5.1.6.2 Potential KPI Requirements 11

5.2 Local AI/ML model split on factory robots 12

5.2.1 Description 12

5.2.6.1 Potential Functionality Requirements 16

5.2.6.2 Potential KPI Requirements 17

6 AI/ML model/data distribution and sharing by leveraging direct device connection 17

6.1 AI Model Transfer Management through Direct Device Connection 17

6.1.1 Description 17

6.1.2 Pre-conditions 18

6.1.3 Service Flows 18

6.1.4 Post-conditions 19

6.1.5 Existing features partly or fully covering the use case functionality 19

6.1.6 Potential New Requirements needed to support the use case 20

6.1.6.1 Potential Functionality Requirements 20

6.1.6.2 Potential KPI Requirements 20

6.2 5GS assisted transfer learning for trajectory prediction 21

6.2.1 Description 21

6.2.2 Pre-conditions 21

6.2.3 Service Flows 22

6.2.4 Post-conditions 22

6.2.5 Existing features partly or fully covering the use case functionality 22

6.2.6 Potential New Requirements needed to support the use case 23

6.2.6.1 Potential Functionality Requirements 23

6.2.6.2 Potential KPI Requirements 23

7 Distributed/Federated Learning by leveraging direct device connection 24

7.1 Direct device connection assisted Federated Learning 24

7.1.1 Description 24

7.1.2 Pre-conditions 25

7.1.3 Service Flows 26

7.1.4 Post-conditions 26

7.1.5 Existing features partly or fully covering the use case functionality 26

7.1.6 Potential New Requirements needed to support the use case 27

7.1.6.1 Functional requirement 27

7.1.6.2 KPI requirement for direct device communication 27

7.2 Asynchronous FL via direct device connection 28

7.2.1 Description 28

7.2.2 Pre-conditions 28

7.2.3 Service Flows 29

7.2.4 Post-conditions 29

7.2.5 Existing features partly or fully covering the use case functionality 29

7.2.6 Potential New Requirements needed to support the use case 30

7.3 5GS assisted distributed joint inference for 3D object detection 30

7.3.1 Description 30

7.3.2 Pre-conditions 31

7.3.3 Service Flows 32

7.3.4 Post-conditions 32

7.3.5 Existing features partly or fully covering the use case functionality 32

7.3.6 Potential New Requirements needed to support the use case 32

7.3.6.1 Potential Functionality Requirements 32

7.3.6.2 Potential KPI Requirements 32

8 Consolidated potential requirements and KPIs 33

8.1 Consolidated potential requirements 33

8.1.1 Enabling 33

8.1.2 QoS support 34

8.1.3 Member selection 35

8.1.4 Charging 35

8.2 Consolidated potential KPIs 36

8.2.1 Split AI/ML operation between AI/ML endpoints 36

8.2.2 AI/ML model/data distribution and sharing by leveraging direct device connection 37

8.2.3 Distributed/Federated Learning by leveraging direct device connection 37

9 Conclusion and recommendations 37

Annex A: Change history 39

# Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

x the first digit:

1 presented to TSG for information;

2 presented to TSG for approval;

3 or greater indicates TSG approved document under change control.

y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.

z the third digit is incremented when editorial only changes have been incorporated in the document.

# Introduction

This document identifies use cases with functional and performance requirements for 5G system support of AI/ML training and inference using direct device connection. By using direct device connection, it can provide a communication capacity and coverage, more work task offloading choice to achieve an efficient AIML training and inference.

# 1 Scope

The objective of this document is to study the use cases with potential functional and performance requirements to support efficient AI/ML operations using direct device connection for various applications e.g. auto-driving, robot remote control, video recognition, etc.

The aspects addressed in the document includes:

- Identify the use cases for distributed AI inference;

- Identify the use cases for distributed/decentralized model training;

- Gap analysis to existing 5GS mechanism to support the distributed AI inference and model training.

# 2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non‑specific.

- For a specific reference, subsequent revisions do not apply.

- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[2] 3GPP TR 22.874, Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS (Release 18)

[3] 3GPP TS 22.104, Service requirements for cyber-physical control applications in vertical domains

[4] Huaijiang Zhu, Manali Sharma, Kai Pfeiffer, Marco Mezzavilla, Jia Shen, Sundeep Rangan, and Ludovic Righetti, “Enabling Remote Whole-body Control with 5G Edge Computing”, to appear, in Proc. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Available at: <https://arxiv.org/pdf/2008.08243.pdf>

[5] B. Kehoe, S. Patil, P. Abbeel, and K. Goldberg, “A survey of research on cloud robotics and automation,” IEEE Transactions on automation science and engineering, vol. 12, no. 2, pp. 398–409, 2015.

[6] M. Chen, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, “Distributed Learning in Wireless Networks: Recent Progress and Future Challenges”IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, VOL. 39, NO. 12, DECEMBER 2021

[7] M. Chen, H. V. Poor, W. Saad, and S. Cui, “Wireless communications for collaborative federated learning,” IEEE Commun. Mag., vol. 58, no. 12, pp. 48–54, Dec. 2020

[8] Jianmin Chen, Xinghao Pan, Rajat Monga, Samy Bengio, Rafal Jozefowicz, “Revisiting Distributed Synchronous SGD,” arXiv preprint arXiv: 1604.00981, 2016

[9] Shuxin Zheng, Qi Meng, Taifang Wang, Wei Chen, Nenghai Yu, Zhi-Ming Ma, Tie-Yan Liu, “Asynchronous Stochastic Gradient Descent with Delay Compression” arXiv: 1609.08326, 2020

[10] 3GPP TR 21.905: "Service requirements for the 5G system".

[11] Yusuf Aytar, Carl Vondrick, Antonio Torralba: "SoundNet: Learning Sound Representations from Unlabeled Video", 27 Oct 2016.

[12] Iacovos Ioannou et al.: "Distributed Artificial Intelligence Solution for D2D Communication in 5G Networks", 20 Jan 2020.

[13] Pimmy Gandotra et al.: "Device-to-Device Communication in Cellular Networks: A Survey".

[14] Davide Villa et al.: "Internet of Robotic Things: Current Technologies, Applications, Challenges and Future Directions", 15 Jan 2021.

[15] Charles R. Qi et al.: "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation", 10 Apr 2017.

[16] 3GPP TS 22.261: "Service requirements for the 5G system".

[17] 3GPP TS 23.303: "Proximity-based services (ProSe); Stage 2".

[18] 3GPP TS 22.104: "Service requirements for cyber-physical control applications in vertical domains; Stage 1".

[19] <https://www.robots.ox.ac.uk/~vgg/software/vgg_face/>.

[20] Y. Kang et al., "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge", ACM SIGPLAN Notices, vol. 52, no. 4, pp. 615–629, 2017.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", in Proc. NIPS, 2012, pp. 1097–1105.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in Proc. IEEE CVPR, Jun. 2016, pp. 770-778.

[23] Zhang Z, Wang S, Hong Y, et al. Distributed dynamic map fusion via federated learning for intelligent networked vehicles[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 953-959.

[24] https://github.com/open-mmlab/OpenPCDet

[25] To Transfer or Not To Transfer, Massachusetts Institute of Technology, MIT, Michael T. Rosenstein, et al.

[26] Wang, J. et al. Easy Transfer Learning by Exploiting Intra-domain Structures. In 2019 IEEE International Conference on Multimedia and Expo (ICME), pages 1210-1215 IEEE.

[27] Wang K C, Fu Y, Li K, et al. Variational model inversion attacks[J]. Advances in Neural Information Processing Systems, 2021, 34: 9706-9719.

[28] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, et. al. Argoverse: 3D Tracking and Forecasting with Rich Maps. arXiv:1911.02620v1 [cs.CV] 6 Nov 2019.

# 3 Definitions, symbols and abbreviations

## 3.1 Definitions

For the purposes of the present document, the terms and definitions given in 3GPP TR 21.905 [1] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in 3GPP TR 21.905 [1].

**Proximity-based work task offloading:** based on 3rd party’s request, a relay UE receives data from a remote UE via direct device connection and performs calculation of a work task for the remote UE. The calculation result can be further sent to network server.

## 3.2 Symbols

For the purposes of the present document, the following symbols apply:

<symbol> <Explanation>

## 3.3 Abbreviations

For the purposes of the present document, the abbreviations given in 3GPP TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in 3GPP TR 21.905 [1].

<ACRONYM> <Explanation>

# 4 Overview

In TR 22.874, three types of AIML operations as below has been described

* AI/ML operation splitting between AI/ML endpoints;
* AI/ML model/data distribution and sharing over 5G system;
* Distributed/Federated Learning over 5G system.

For the phase-2 study, it continues to study how the 5GS can have more gains for each type of AIML operations when leveraging direct device connection. Thus, the following clause 5, 6, and 7 is to capture use cases corresponding to the three types of AIML operations considering the assistance of direct device connection.

# 5 Split AI/ML operation between AI/ML endpoints for AI inference by leveraging direct device connection

## 5.1 Proximity based work task offloading for AI/ML inference

### 5.1.1 Description

The model splitting is the most significant feature for AI inference. As some R18 use cases in TR 22.874[2] shows, the number of terminal computing layers and the amount of data transmission are corresponding to different model splitting points. For example, as figure 5.1-1 shows, the general trend is that the more layers the UE calculated, the less intermediate data needs to be transmitted to application server. In another word, when UE has low computation capacity (e.g. due to low battery), the application can change the splitting point to let UE calculate fewer layers while increasing the data rate in Uu for transmitting a higher load of intermediate data to network.

However, sometimes the data rate cannot be increased due to radio resource limitation, in such circumstances, UE with low computation capacity needs to offload the computation task to a proximity UE (likely a relay UE) but still keeping the computation service and let the proximity UE to send the calculated data to network. Thus, by offloading the work task using direct device connection, the original UE’s computation load will be released while the data rate in Uu interface will not necessarily be increased either, which leads to a more ideal performance.

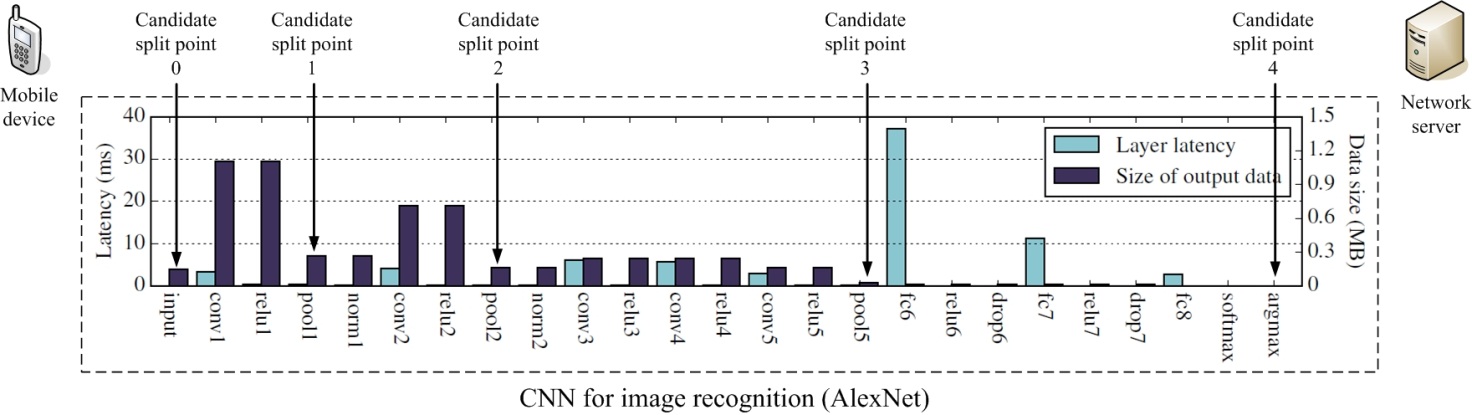


Figure 5.1-1. Layer-level computation/communication resource evaluation for an AlexNet model (abstracted from subclause 5.1.1 in TR 22.874)

### 5.1.2 Pre-conditions

A UE uses the AI model (AlexNet) for image recognition. As predetermined by application, there are 5 alternative splitting points which are corresponding to intermediate data size and data rate, see reference [13-14] in TR 22.874, while fewer the layers being calculated implies fewer the workload being performed by UE. The specific values are shown in the table below (it is abstracted from clause 5.1 Split AI/ML image recognition in TR22.874).

Table 5.1-1: Required UL data rate for different split points of AlexNet model for video recognition @30FPS (Frame Per Second)

|  |  |  |
| --- | --- | --- |
| **Split point** | **Approximate output data size (MByte)** | **Required UL data rate (Mbit/s)** |
| Candidate split point 0  (Cloud-based inference) | 0.15 | 36 |
| Candidate split point 1  (after pool1 layer) | 0.27 | 65 |
| Candidate split point 2  (after pool2 layer) | 0.17 | 41 |
| Candidate split point 3  (after pool5 layer) | 0.02 | 4.8 |
| Candidate split point 4  (Device-based inference) | N/A | N/A |

### 5.1.3 Service Flows

(a) no task offloading (b) task offloading by UE-B

Figure 5.1-2: Using direct device connection (sidelink) to realize the proximity-based work task offloading. In this case, the data rate on Uu need not be increased while the original UE’s computation load is offloaded

1. As shown in left(a) of Figure 5.1-2, UE-A is doing image recognition using Alexnet Model as described in clause 5.4.2. It selects splitting point-3 for the AI inference.

The E2E service latency (including image recognition latency and intermediate data transmission latency) is 1 second.

1. When the UE-A’s battery becomes low, it cannot afford the heavy work task for the AlexNet model (i.e. calculating layer 1-15 for AlexNet model in local side).
2. Being managed by 5G network, the UE-A discovers UE-B (a Customer Premise Equipment, CPE) which has installed the same model and is willing to take the offloading task from UE-A.

NOTE 1: The 5G network does not store UE-A and UE-B’s location data.

Then UE-A established the sidelink (direct device connection) to UE-B. During the sidelink establishment, the UE-B also gets the information of the total service latency (including the image recognition latency and intermediate data transmission latency) and the processing time consumed by UE-A for computing layer 1-4.

Since the UE-B has acquired the E2E service latency and the processing time consumed by UE-A, and also it knows its own processing time for computing layer 5-15, the UE-B can determine the QoS parameters applied to both Uu and Sidelink while keeping the E2E service latency same as the E2E service latency described in step-1.

NOTE 2: It is assumed that the UE-A and UE-B have the same computation capacity, i.e. the time used for computing the certain AlexNet model layers are the same for UE-A and UE-B. Otherwise, the data rate on Uu and Sidelink may be changed accordingly.

1. The UE-A sends the intermediate data (data after calculating layer 1-4) to UE-B via sidelink and let UE-B make further processing then transmit the intermediate data (data after calculating layer 5-15) to application server via Uu. The specific model layers being computed by UE-A and UE-B are shown in the right(b) in figure 5.1-2.
2. UE-A continues to perform image recognition by leveraging sidelink and UE-B’s computation capacity while the source and destination IP address and the E2E service latency for the image recognition service is unchanged.

### 5.1.4 Post-conditions

Thanks to UE-B’s help, the proximity-based work task offloading is performed. By doing so,

- it decreased the UE-A’s work task by letting UE-A to compute fewer layers of AlexNet model, which helps to meet the low battery condition happened to UE-A;

- the UE-B computes the rest of layers which is originally from the UE-A’s work task;

- the mobile network does not need to increase the QoS parameters such as guaranteed data rate because the intermediate data rate transmitted by UE-B is unchanged.

### 5.1.5 Existing features partly or fully covering the use case functionality

In TS 22.261 clause 6.9, the description about the direct network connection mode and the indirect network connection mode as well as the service continuity for switching between the two modes have been described. They are summarized as below:

*The UE (remote UE) can connect to the network directly (direct network connection), connect using another UE as a relay UE (indirect network connection), or connect using both direct and indirect connections.*

*The 5G system shall support different traffic flows of a remote UE to be relayed via different indirect network connection paths.*

*The 5G system shall be able to maintain service continuity of indirect network connection for a remote UE when the communication path to the network changes (i.e. change of one or more of the relay UEs, change of the gNB).*

However, there is no proximity-based work task offloading which means that the “relay UE” not only performs the indirect network communication but also performs task computation for the “remote UE”. This may impact the current discovery mechanism, QoS determination on Uu and PC5, and charging aspect.

### 5.1.6 Potential New Requirements needed to support the use case

#### 5.1.6.1 Potential Functionality Requirements

[P.R.5.1.6-001] The 5G system shall be able to support the means to modify the communication QoS ensuring the end-to-end latency can be satisfied when a relay UE is involved for a proximity-based work task offloading.

NOTE 1: Due to the proximity-based work task offloading, the data size transmitted via sidelink and Uu of the indirect network connection is different

[P.R.5.1.6-002] The 5G system shall be able to collect charging information for proximity-based work task offloading.

[PR.5.1.6-003] The 5G system shall support service continuity when a UE communication path changes between a direct network connection and an indirect network connection, including the case when the data size transmitted over the two connection is different (e.g. for a proximity-based work task offloading).

#### 5.1.6.2 Potential KPI Requirements

Considering the widely-used AlexNet and VGG-16 model for proximity-based work task offloading, the following KPIs need to be supported:

Table 5.1-2 KPI requirements for proximity-based work task offloading

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **UL data size**  **(for sidelink)** | **UL data rate**  **(for sidelink)** | **Intermediate data uploading latency (including sidelink+Uu)** | **Image recognition latency** |
| AlexNet model with 30FPS (NOTE 1) | 0.15 - 0.02Mbyte for each frame | 4.8 – 65 Mbit/s | - 2ms for Remote driving, AR displaying/gaming, and remote-controlled robotics;  - 10ms for video recognition;  - 100ms for One-shot object recognition, Person identification, or photo enhancement in smart phone | 1s |
| VGG-16 model with 30FPS | * 1. - 1.5 Mbyte for each frame | 24 - 720 Mbit/s | 1s |
| **NOTE 1: FPS stands for Frame Per Second** | | | | |

## 5.2 Local AI/ML model split on factory robots

### 5.2.1 Description

In a modern factory, a team on a workstation comprises two human operators, two mobile robots and a fixed robot. Everyone has his own pre-defined task. Robots assist the human operators by accomplishing painful tasks in a fluid and precise manner; they also monitor that the workstation environment remains safe for the human operators. The mobile robots must not interfere with humans and between them.

The robot control is not executed on a distant server in the cloud because reliability and confidentiality were not ensured at a sufficient level. Furthermore, as stated in [14], the overall end-to-end latency is not always guaranteed which can cause production loss. Communications between robots can rely on private wireless networks in the factory that enable expected QoS (reliability, throughput, and latency) as well as confidentiality.

The new robots are autonomous robots that can react to human voices or learn in real-time what operators do. They can perceive their environment and transmit information to other robots. They can communicate, learn from each other, assist each other and do self-monitoring.

The autonomous robot’s skills rely on several AI/ML models running on the robot itself which has the inconvenience that the mobile robot’s battery drain quicker. To overcome this issue, when the battery level reaches a certain value, a part of the AI/ML model can be transferred to a service hosting environment and / or to another robot by splitting the AI/ML model as defined in [20]. The split model approach of [20] is applicable to a UE-to-UE (or robot-to-robot) architecture. Thus, the AI/ML model M is split and shared between (e.g.) 2 robots, say an assisted robot and an assistant robot. Intermediate data generated by the assisted robot are transferred to the assisting robot which finalizes the inference and transmits the results back to the assisted robot. This intermediate data transfer must be extremely efficient in terms of latency and throughput. When many models are at stake, the split model method is an additional challenge for the 5GS in terms of throughput, latency and synchronization.

Because they are more autonomous, mobile and smart, the industry 4.0 robots embed a large variety of sensors that generate huge amount of data to process. Table 5.2.1-1 reflects that variety.

Each type of sensing data requires a different AI/ML model. Each of these models produce predictions in a certain delay and with a certain accuracy.

Thus, as an offloading strategy, we can imagine that a model is split between 2 robots because it has been established that for a particular AI/ML model, the latency with the sidelink communication was better (smaller) than with the regular 5G communication path, as stated in [12] and [13].

At the same time, other AI/ML models are split between the robot and the service hosting environment because from an energy standpoint this configuration is the best.

Table 5.2.1-2 is an example of this offloading strategy where four AI/ML models are split between a robot and the service hosting environment and four other AI/ML models are split between two robots.

Sidelink and 5G communication paths are complementary from an AI/ML model split policy standpoint.

Table 5.2.1-1 shows some typical and diverse AI/ML models that can be used on robots. For each model, all the split point candidates have been considered and only the split points that generate the minimum and the maximum amount of intermediate data have been noted.

Table 5.2.1-1: Intermediate AI/ML data size per AI/ML model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Name** | **Model type** | **Intermediate data size (MB)** | | | |
| **8 bits data format** | | **32 bits data format** | |
| **Min** | Max | Min | **Max** |
| AlexNet [21] | Image recognition | **0.02** | 0.06 | 0.08 | **0.27** |
| ResNet50 [22] | Image recognition | **0.002** | 1.6 | 0.008 | **6.4** |
| SoundNet [11] | Sound recognition | **0.0017** | 0.22 | 0.0068 | **0.88** |
| PointNet [15] | Point Cloud | **0.262** | 1.04 | 0.0068 | **4.19** |
| VGGFace [19] | Face recognition | **0.000016** | 0.8 | 0.000064 | **3.2** |
| Inception resnet | Face recognition | **0.0017** | 0.37 | 0.0068 | **1.51** |

In Table 5.2.1-2 the AI/ML models are distributed between the service hosting environment and the proximity robot. The way the models are distributed is out of scope of this use case and depends on various criteria as said previously. Therefore, the next table is an example that illustrates the distribution. The intermediate data size is presented with the range [Min – Max], where Min and Max are respectively the sum of the Min values and the sum of the Max values of the selected models as defined in Table 5.2.1-1 (figures in bold).

Table 5.2.1-2: Example of models distribution and data rate for intermediate AI/ML data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Name** | **Offloading target** | **Intermediate data size (MB)** | **Transfer time (ms)** | **Data rate (Gb/s)** |
| AlexNet [21] | Proximity robot or Service Hosting Environment | [0.000016 – 1.6]  (8 bits data format) | 10 | [0.128 - 1.28] |
| ResNet50 [22] |
| VGGFace [19] |
| SoundNet [11] | [0.000064 – 6.4]  (32 bits data format) | 10 | [0.512 – 5.12] |
| PointNet [15] |
| Inception resnet |
|  |

As previously said latency is a critical requirement. Figure 5.2.1-3 summarizes the latency cost in three scenarios:

1. The inference of model M is done locally. Latency is denoted LLI.
2. The inference process is fully offloaded on a second device (Robot/UE). Latency is denoted LFO.
3. The inference process is partially offloaded on a second device (Robot/UE). Latency is denoted LPO.

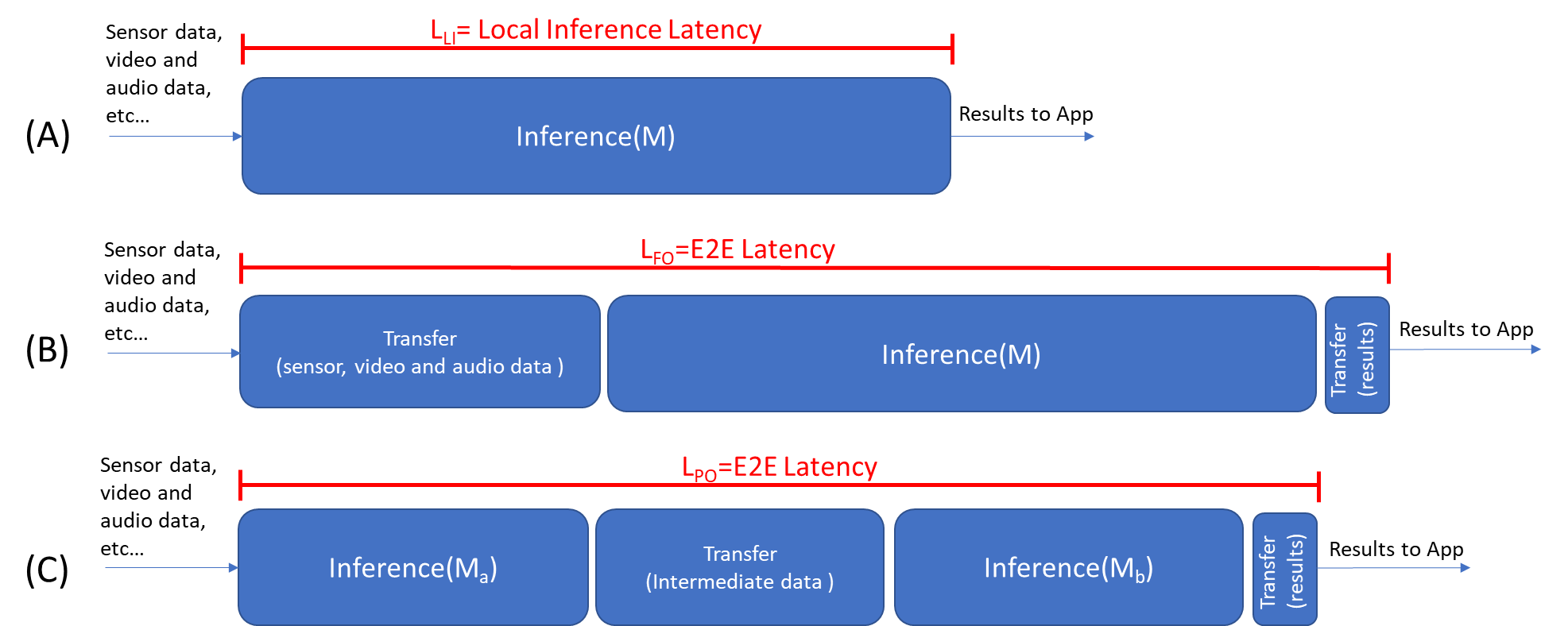


Figure 5.2.1‑3 Latency summary

The current Use Case promotes the scenario (C) where a model M is split in two sub-models Ma and Mb. If both robots (UEs) have a similar computing power, the assumption is that the latency due to the inference of model M is almost equal to the latency of model Ma plus the latency of model Mb.

Hence, once the split model is deployed on the two robots (UEs), the aim is to minimize the E2E latency and to be as close as possible to the non-split case. This is done with a transfer delay of both the intermediate data and the inference results as small as possible. We can note that if the computing power on the assistant robot is more important, then scenario (C) would be the preferred scenario.

In scenario (B), the inference process is fully offloaded on the assistant robot (UE). The major inconvenience is the strong and negative impact on latency of the raw data transfer towards the assistant robot.

5.2.2 Pre-conditions

Two human operators are working.

Two mobile AI-driven robots (Arobot and Brobot) and one static AI-driven robot (Crobot) assist them.

The three robots (Arobot, Brobot and Crobot) belong to the same service area, embed the same two powerful AI/ML models M1 and M2, sensors (e.g. LIDAR, microphone) and cameras (e.g., 8K videos stream).

Arobot and Brobot are powered with a battery and Crobot with fixed ground power.

The three robots (Arobot, Brobot and Crobot) are connected, e.g., to the AF, 5GC, or to each other using D2D technologies (Prose, BT, WiFi, etc.).

The workstation is equipped with camera and sensors.

The service area is 30 m x 30 m and the robot speed is at maximum 10 km/h.

The service area is covered by a small cell and a service hosting environment is connected and can support AI/ML processes.

5.2.3 Service Flows

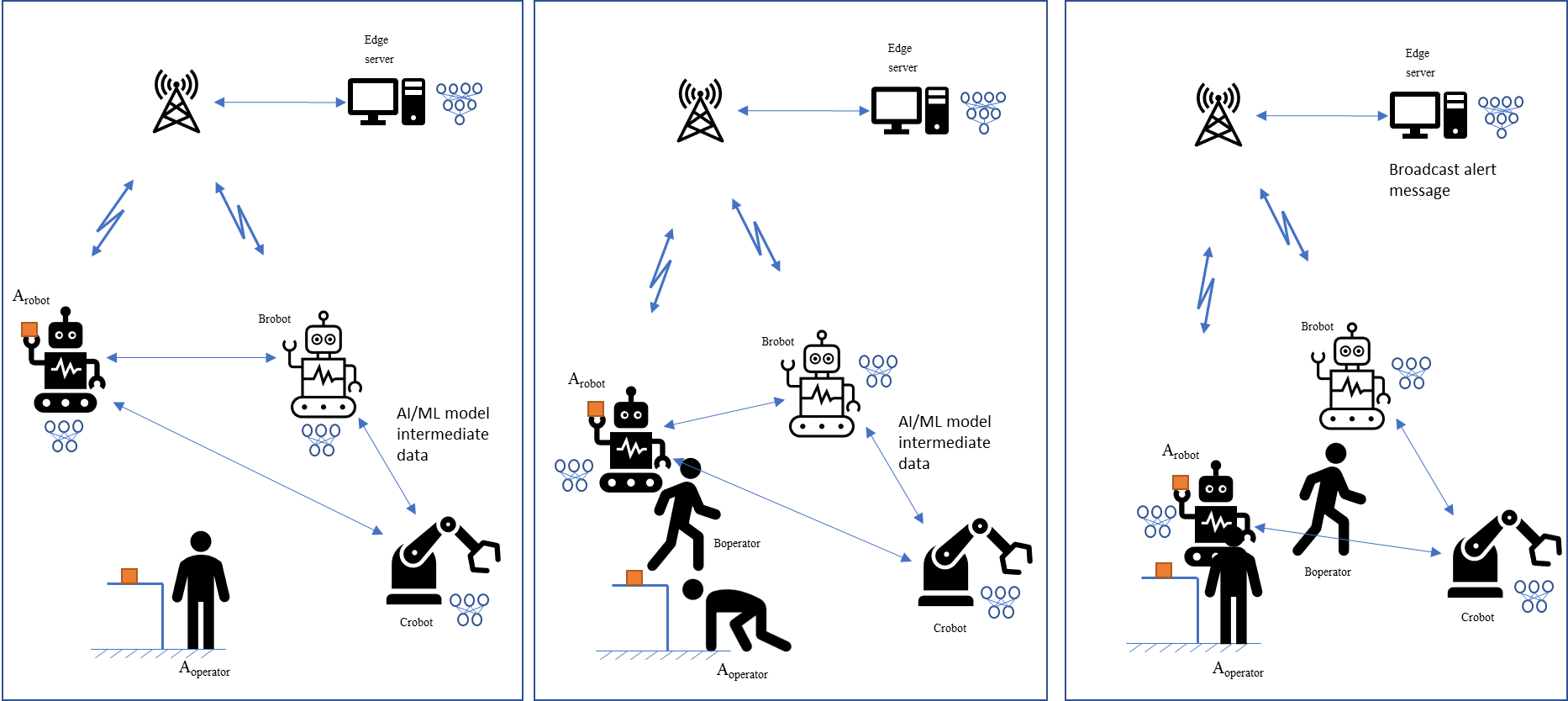


Figure 5.2.3‑1 Factory service flow

1. Brobot battery level is rather low but it can still work for a while if a part of its machine learning process is offloaded.
2. Brobot broadcasts a request message to get assistance. Crobot and the service hosting environment responds positively.
3. Brobot negotiates with Crobot and the service hosting environment what parts of M1 and M2 respectively for the inference process they are in charge of, knowing that the quality of the prediction must not be under a certain level and that the end-to-end latency must not be above a certain value. M1 model is split between Brobot and the service hosting environment. M2 model is split between Brobot and Crobot.
4. Brobot, Crobot and the service hosting environment agree on split points for both M1 and M2 models and Brobot starts sending the intermediate data to Crobot and the service hosting environment.
5. Crobot infers and transmits using unicast mode with a very short delay the predictions back to Brobot. The service hosting environment infers and transmits using unicast mode the predictions back to Brobot in a very short delay.
6. In the meanwhile, Arobot is carrying a load to the operator Aoperator.
7. Aoperator bends down to pick up a screw that has fallen on the floor. At the same time Boperator is passing between Aoperator and Arobot. Arobot can’t see Aoperator anymore.
8. Brobot is busy with another task, but it can observe the scene. It reports the scene as intermediate data to Crobot and the service hosting environment.
9. Crobot and the service hosting environment amend the ML model based on the new training data.
10. Crobot and the service hosting environment infer and then transmit in unicast the prediction back to Brobot. The safety application on the service hosting environment collects the inference results.

5.2.4 Post-conditions

Intermediate data can be exchanged between two robots (UEs) and / or service hosting environment, and the robot with a low battery level can continue working for a while.

All the robots in the group receive the alert message and react:

1. They all stop working; or
2. Arobot changes its trajectory.

Aoperator and Boperator can work safely.

The huge amount of data that is required for inferring is kept local in the factory.

5.2.5 Existing features partly or fully covering the use case functionality

The Use Case can rely on the Proximity Service (ProSe) services as defined in 3GPP TS 23.303 [17].

Cyber-Physical Control Applications, see 3GPP TS 22.104 [18], already proposes to rely on a ProSe communication path. The proposed requirements are limited in terms of data transfer as shown in Table 5.2-1, where the message size does not exceed a few hundred of Bytes (250 kB at maximum).

3GPP TS 22.261 [16] clause 6.40 provides requirements for AI/ML model transfer in 5GS. The requirements in this clause does not consider requirements for direct device connection.

In 3GPP TS 22.261 [16] Table 7.6.1-1, the max. end-to-end latency is 10 ms, the maximum data rate is [1] Gbits/s and reliability is 99.99% for Gaming or Interactive Data Exchanging.

Table 7.6.1-1 KPI Table for additional high data rate and low latency service

| Use Cases | Characteristic parameter (KPI) | | | Influence quantity | | |
| --- | --- | --- | --- | --- | --- | --- |
| Max allowed end-to-end latency | Service bit rate: user-experienced data rate | Reliability | # of UEs | UE Speed | Service Area  (note 2) |
| Gaming or Interactive Data Exchanging  (note 3) | 10ms (note 4) | 0,1 to [1] Gbit/s supporting visual content (e.g. VR based or high definition video) with 4K, 8K resolution and up to120 frames per second content. | 99,99 % (note 4) | ≤ [10] | Stationary or Pedestrian | 20 m x 10 m; in one vehicle (up to 120 km/h) and in one train (up to 500 km/h) |
| NOTE 1: Unless otherwise specified, all communication via wireless link is between UEs and network node (UE to network node and/or network node to UE) rather than direct wireless links (UE to UE).  NOTE 2: Length x width (x height).  NOTE 3: Communication includes direct wireless links (UE to UE).  NOTE 4: Latency and reliability KPIs can vary based on specific use case/architecture, e.g. for cloud/edge/split rendering, and can be represented by a range of values.  NOTE 5: The decoding capability in the VR headset and the encoding/decoding complexity/time of the stream will set the required bit rate and latency over the direct wireless link between the tethered VR headset and its connected UE, bit rate from 100 Mbit/s to [10] Gbit/s and latency from 5 ms to 10 ms.  NOTE 6: The performance requirement is valid for the direct wireless link between the tethered VR headset and its connected UE. | | | | | | |

These requirements partially cover the current Use Case needs.

5.2.6 Potential New Requirements needed to support the use case

#### 5.2.6.1 Potential Functionality Requirements

[P.R.5.2.6-001] Subject to user consent and operator policy, the 5G system shall support the transfer of AI/ML model intermediate data from UE to UE via the direct device connection.

[P.R.5.2.6-002] Subject to user consent and operator policy, the 5G system shall be able to provide means to predict and expose network condition changes (i.e. bitrate, latency, reliability) and receive user preferences on usage of the direct device connection or the direct network connection in order to meet the user experienced data rate and latency.

[P.R.5.2.6-003] Subject to user consent and operator policy, the 5G system shall be able to dynamically select the intermediate device that is capable to perform the needed functionalities, e.g., AIML splitting.

[P.R.5.2.6-004] Subject to user consent and operator policy, the 5G system shall be able to maintain the QoS (latency, reliability, data rate as defined in the Table 5.2.6.2-1 below) of the communication path of the direct device connection.

[P.R.5.2.6-005] Subject to user consent and operator policy, the 5G system shall be able to have the means to modify the QoS of the communication path of the direct device connection.

NOTE: The split point selection is dynamic. In consequence, the amount of intermediate data will vary. To maintain the QoS, the bandwidth is adjusted.

#### 5.2.6.2 Potential KPI Requirements

Based on Table 5.2.1-2, the potential KPI requirement is as below:

Table 5.2.6.2-1 KPI for intermediate AI/ML data transmission for model split based robot control

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model Name** | **Payload size (Intermediate data size)** | **Max allowed end-to-end latency** | **Experienced data rate** | **Service area dimension** | **Communication service availability** | **Reliability** |
| AlexNet [21] | 0.000016 – 1.6 MByte  (8 bits data format) | 10 ms | 0.128 - 1.28 Gbps | 900 m2 (30 m x 30 m) | 99.999 % | 99.999 % |
| ResNet50 [22] |
| VGGFace [19] |
| SoundNet [11] |
| PointNet [15] | 0.000064 – 6.4 Mbyte  (32 bits data format) | 10 ms | 0.512 - 5.12 Gbps |
| Inception resnet |

# 6 AI/ML model/data distribution and sharing by leveraging direct device connection

## 6.1 AI Model Transfer Management through Direct Device Connection

### 6.1.1 Description

Based on the earlier study in phase one 3GPP TR 22.874 [2], operators can provide services to help manage and distribute the AI/ML models especially in the edge server so that the UE can acquire a proper model immediately. However, when a lot of UEs requesting for the same model at the same time or the UE is blocked by barriers with poor connection with the base station, the model transfer process will become longer than expected.

To overcome this difficulty, as shown in Fig.1, a volunteer UE which is well connected to the base station can help relaying AI/ML models or receive and store AI/ML models first. Then, the other UEs can download AI/ML models from the volunteer UE through direct device connection. In this way, all UE can have a stable and reliable model transfer process while the radio resource of the base station can be saved. Besides, the volunteer UE can transfer the stored models to other volunteer UEs under operator’s control.

The selection of volunteer UE can be realized by local network policies and strategies. And it also can be exposed as a capability to the 3rd party company when the company wants to choose one or a few certain UEs to be volunteer UEs in an activity. For example, a travel company may assign the tour guides’ Augmented Reality (AR) headsets as volunteer UEs in a carnival through the operator’s network exposure. The travel company may sign a higher quality plan for tourist guides’ devices to provide better user experience for following tourists. Meanwhile, operator can benefit from the alternative open service based on AI/ML model management capabilities and may avoid low Quality of Service due to crowding direct connections to base stations during the carnival.

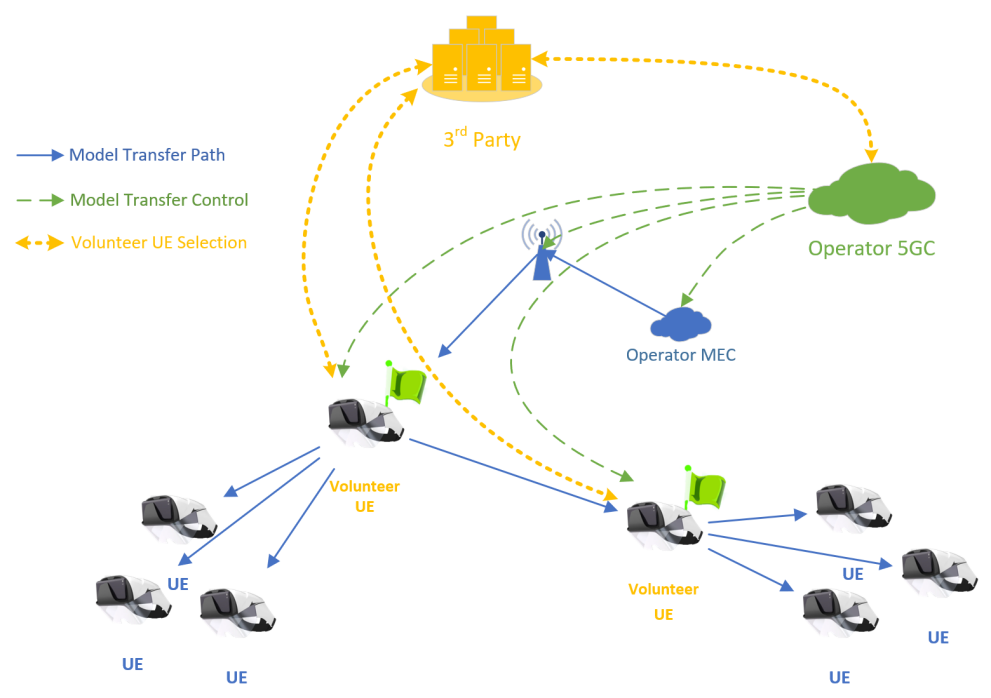


Fig.1 AI/ML Model management through direct device connection

### 6.1.2 Pre-conditions

The operator’s MEC near the Jurassic Park stores a variety of AI/ML models according to the the park company’s requirements. And it is capable to transfer the stored model to the device such as AR headset.

The operator rolls out a new high-quality plan which can allow the user customizes own Service Level Agreement (SLA) for specific network address access and data (e.g. AI/ML Models) download. As a trade-off, the user’s device will help transfer the same data through direct device connection to nearby devices sharing common aspiration.

The AR headset can transfer the stored AI/ML model to the other AR headsets. However, the AR headset cannot store all models for different scenarios due to limited storage. Indeed, a model needs to be downloaded when or a few seconds before the UE first appears in the certain area.

Alice and Bob are tour guide hired by Jurassic Park and their real-time positions can be acquired when they are in the park based on signed agreements.

All of AR headsets should in the coverage of the base stations serving for the Jurassic Park.

### 6.1.3 Service Flows

1. Jurassic Park provides panorama AR tour guide services in a commercial area and a tropical rainforest area. AR headsets need to download Model A and B (both are VGG-16, 552MByte) respectively.

2. To provide high quality user experience, Jurassic Park company indicates to the operator that AR headsets require to download model A in area A and model B in area B.

3. The Jurassic Park company signs a high-quality plan for tour guide Alice and Bob’s AR Headsets for providing better service to the tour group using direct device connection.

4. When Bob and his tour group enter area A, their headsets request for the Model A. The operator network finds they requested the same model and Bob is a signed volunteer UE, then triggers to establish a QoS acceleration for Bob’s model downloading timely within 1 second. Meanwhile, Jurassic Park requests the operator network to inform Bob to help transfer the model to all other UE near Bob. Also, the operator network informs all other UE near Bob that Bob can provide the model as well. The UE which is a little far from Bob (e.g. out of Bob’s coverage) will still download the model through the base station directly.

5. Alice and her group are 10 meters far from Bob and also move towards to area A. Jurassic Park predicts their desire model based on their movement and finds Bob has already downloaded it based on the model transferring records. Jurassic Park requests operator network to inform Alice that she can request model from Bob. Meanwhile, the operator network indicates all other UE near Alice to download the model from Alice.

6. For Alice and Bob, they can see the status of all direct device connections to themselves through network exposures providing by operators (e.g. monitored bandwidth and latency of each direct device connection)

7. When Alice and Bob notice that their groups have a poor QoS of model transfer through direct device connection, they can send a request to the park company for promoting the performance of their direct device connections and the park company will send a similar message to the operator through network exposure to active a temporary acceleration of these direct device connections (e.g. expand the bandwidth of each direct device connection).

### 6.1.4 Post-conditions

1. The tourists can enjoy the continuous AR services with smooth model switchover when their location and responding models change.

2. Tour group’s AR headsets provides user experience of the panorama AR tour guide services that can help retrain and improve AI/ML models in operator’s MEC by Jurassic Park company (e.g. Federated/Distributed Learning).

3. the operator network performs analytics, based on network statistics and quality of experience reported by Jurassic Park company, to improve and optimized the model transfer process (e.g. setting constraints for maximum direct device connection for one volunteer UE and choose a temporary volunteer UE for sharing model transfer task).

### 6.1.5 Existing features partly or fully covering the use case functionality

**In 3GPP TS 22.261 [8] clause 6.27.2 "Requirements"**

The 5G system shall be able to make the position-related data available to an application or to an application server existing within the PLMN, external to the PLMN, or in the User Equipment.

**In 3GPP TS 22.261 [8] clause 6.9.2.4 "Relay UE Selection"**

The 3GPP system shall support selection and reselection of relay UEs based on a combination of different criteria e.g.

- the characteristics of the traffic that is intended to be relayed (e.g. expected message frequency and required QoS),

- the subscriptions of relay UEs and remote UE,

- the capabilities/capacity/coverage when using the relay UE,

- the QoS that is achievable by selecting the relay UE,

- the power consumption required by relay UE and remote UE,

- the pre-paired relay UE,

- the 3GPP or non-3GPP access the relay UE uses to connect to the network,

- the 3GPP network the relay UE connects to (either directly or indirectly),

- the overall optimization of the power consumption/performance of the 3GPP system, or

- battery capabilities and battery lifetime of the relay UE and the remote UE.

NOTE: Reselection may be triggered by any dynamic change in the selection criteria, e.g. by the battery of a relay UE getting depleted, a new relay capable UE getting in range, a remote UEs requesting additional resources or higher QoS, etc.

**In 3GPP TS 22.261 [8] v18.6.0 clause 6.40.2**

Based on operator policy, the 5G system shall be able to provide an indication about a planned change of bitrate, latency, or reliability for a QoS flow to an authorized 3rd party so that the 3rd party AI/ML application is able to adjust the application layer behaviour if time allows. The indication shall provide the anticipated time and location of the change, as well as the target QoS parameters.

Subject to user consent, operator policy and regulatory constraints, the 5G system shall be able to support a mechanism to expose monitoring and status information of an AI-ML session to a 3rd party AI/ML application.

NOTE: Such mechanism is needed for AI/ML application to determine an in-time transfer of AI/ML model.

Subject to user consent, operator policy and regulatory requirements, the 5G system shall be able to expose information (e.g. candidate UEs) to an authorized 3rd party to assist the 3rd party to determine member(s) of a group of UEs (e.g. UEs of a FL group).

### 6.1.6 Potential New Requirements needed to support the use case

#### 6.1.6.1 Potential Functionality Requirements

[P.R.6.1-001] Subject to user consent, operator policies and regional or national regulatory requirements, the 5G system shall be able to support means to monitor a direct device connection and expose corresponding monitoring information (e.g. experienced data rate, latency) to an authorized 3rd party.

NOTE: The monitoring information in [P.R.6.1-001] doesn’t include any user position-related data.

[P.R.6.1-002] Subject to user consent and operator policies, the 5G system shall be able to provide means for an authorized third-party to authorize a group of UEs to exchanging data with each other via direct device connection.

[P.R.6.1-003] The 5G system shall support a mechanism for an authorized third-party to negotiate a suitable QoS of direct device connections for a group of UEs to exchange data with each other.

[P.R.6.1-004] Subject to user consent, operator policies and regulatory requirements, the 5G system shall support means to monitor, characteristics of traffic relayed by a UE participating in the communication and expose to 3rd party.

#### 6.1.6.2 Potential KPI Requirements

[P.R.6.1-005] The 5G system shall support to use direct device communication to transmit the AI/ML model of image recognition and 3D object recognition with the following KPIs.

Table 6.1.6.2-1: KPIs for image recognition and 3D object recognition using direct device connection

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Type** | **Max allowed DL end-to-end latency** | **Experienced data rate**  **in PC5** | **Model size** | **Communication service availability** |
| AlexNet | 1s | 1.92 Gbit/s | 240 MByte | 99.9 % |
| ResNet-152 | 1s | 1.92 Gbit/s | 240 Mbyte | 99.9 % |
| ResNet-50 | 1s | 0.8 Gbit/s | 100 Mbyte | 99.9 % |
| GoogleNet | 1s | 0.218 Gbit/s | 27.2 Mbyte | 99.9 % |
| Inception-V3 | 1s | 0.736 Gbit/s | 92 Mbyte | 99.9 % |
| PV-RCNN | 1s | 0.4 Gbit/s | 50 Mbyte | 99.9 % |
| PointPillar | 1s | 0.14 Gbit/s | 18 Mbyte | 99.9 % |
| SECOND | 1s | 0.16 Gbit/s | 20 Mbyte | 99.9 % |
| For the size of image recognition model, it refers to table 6.1.1-1 in TR22.874 [2], for the size of 3D object recognition model, see [24].  Reliability is assumed to be [99.9 – 99.999]% | | | | |

## 6.2 5GS assisted transfer learning for trajectory prediction

### 6.2.1 Description

AIML model transfer learning is beneficial for lowing cost and raising effective when training a model using a target UE based on a pre-training model. The principle of transfer learning is to use the knowledge from the source domain to train a model in the target domain to achieve more expedient and higher accuracy efficiency [25].

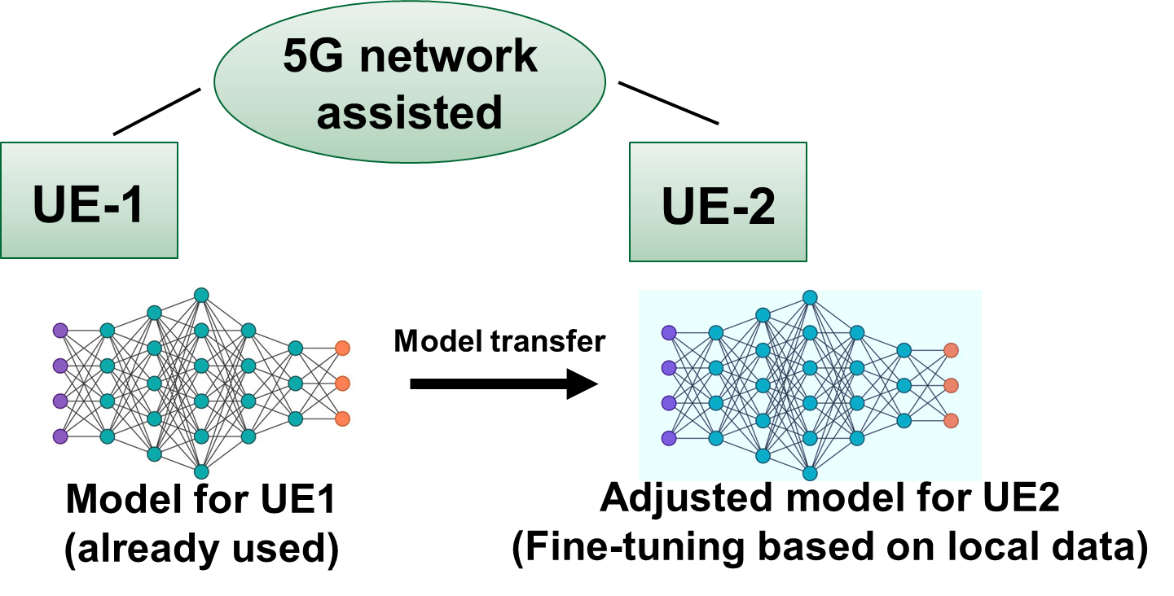


Figure 6.2-1 AI/ML model transfer learning from source UE to target UE [26]

Since the AI model is a kind of knowledge, when the centralized application server acquires enough number of AIML model used by UEs, it may perform a backward inference/inversion attacks [27] to derive the feature of UE’s local data set, which means a privacy risk exists. In order to resolve the privacy concern for transfer learning, the model transfer via direct device connection is a better to be used so that the network node (e.g. application server) cannot acquire the AIML model used by UE and no way to do backward inference.

### 6.2.2 Pre-conditions

Alice is a customer of intelligent-driving service provided by company-A. She lives in Chaoyang district in Beijing and driving to her office building in CBD every working day. By using the intelligent driving service, Alice’s car can predict the trajectory of neighbouring vehicles (as figure 6.2.2 shows), so as to pre-alert Alice of some potential collision and Alice can decide whether to steer, accelerate, or any other driving operation.

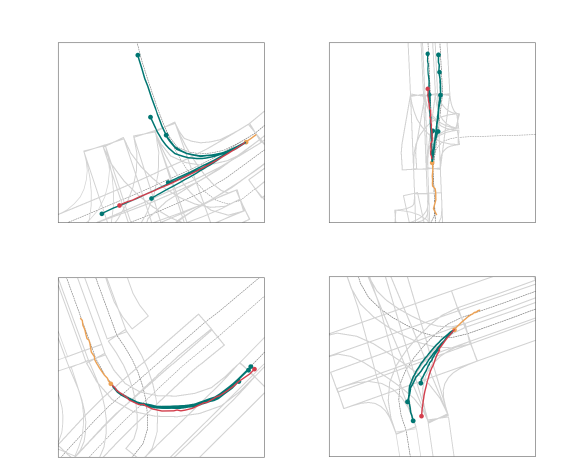


Figure 6.2-2: Qualitative results using model of trajectory prediction: the orange trajectory represents the observed 2s. Red represents ground truth for the next 3 seconds and green represents the multiple forecasted trajectories for those 3s [24].

An AIML model can be for the object recognition and prediction, the model is offered by company-A and customers of company-A have signed “smart driving project” (an agreement for AIML model sharing and improvement).

### 6.2.3 Service Flows

1. Bob bought a car equipped with intelligent driving functionality and he would like to use auto-driving for his daily driving, so he applies to company-A to offer the intelligent-driving service.
2. Company-A needs to install certain AIML model to Bob’s car while use Bob’s local data to train the model. The company-A identified Alice’s model to be shared to Bob’s car.

In order to minimize privacy issue, the “smart driving project” signed by customer only allows the model to be transferred among users directly instead of letting application server to acquire and forward it.

1. Company-A requests 5G system to transmit the AIML model for intelligent driving from Alice’s car to Bob’s car via direct device connection at a proper time (e.g. when the direct device connection can be established)
2. When acquiring the AI model from Alice’s car, Bob’s car performs “fine-tuning” operation of transfer learning based on the local data to tune the model to be better used for its own intelligent-driving service.

### 6.2.4 Post-conditions

Thanks to 5GS assisted AIML model transfer via direct device connection, Bob’s car efficiently gets an ideal AIML model for intelligent-driving by means of transfer learning.

### 6.2.5 Existing features partly or fully covering the use case functionality

**In 3GPP TS 22.261 [8] v18.6.1 clause 6.9**

The 5G system shall support different traffic flows of a remote UE to be relayed via different indirect network connection paths.

The connection between a remote UE and a relay UE shall be able to use 3GPP RAT or non-3GPP RAT and use licensed or unlicensed band.

The connection between a remote UE and a relay UE shall be able to use fixed broadband technology.

The 5G system shall be able to provide indication to a remote UE (alternatively, an authorized user) on the quality of currently available indirect network connection paths.

The 5G system shall be able to maintain service continuity of indirect network connection for a remote UE when the communication path to the network changes (i.e. change of one or more of the relay UEs, change of the gNB).

The 5G system shall be able to support a UE using simultaneous indirect and direct network connection mode.

The 5G system shall enable the network operator to authorize a UE to use indirect network connection. The authorization shall be able to be restricted to using only relay UEs belonging to the same network operator. The authorization shall be able to be restricted to only relay UEs belonging to the same application layer group.

**In 3GPP TS 22.261 [8] v18.6.1 6.40**

Subject to user consent, operator policy and regulatory requirements, the 5G system shall be able to expose information (e.g. candidate UEs) to an authorized 3rd party to assist the 3rd party to determine member(s) of a group of UEs (e.g. UEs of a FL group).

### 6.2.6 Potential New Requirements needed to support the use case

#### 6.2.6.1 Potential Functionality Requirements

[P.R.6.2-001] Based on user consent and 3rd party request, operator policy, the 5G system shall support a means to authorize specific UEs to transmit data (e.g. AI-ML model tansfer for a specific application) via direct device connection in a certain location and time.

[P.R.6.2-002] Subject to user consent and operator policy, the 5G system shall be able to expose information to an authorized 3rd party to assist the 3rd party to determine candidate UEs for data transmission via direct device connection (e.g. for AIML model transfer).

#### 6.2.6.2 Potential KPI Requirements

[P.R.6.2-003] The 5G system shall be able to support transmitting an AI/ML model via direct device connection fulfilling the KPIs for transmission of typical AIML model for trajectory prediction and object recognition [24][28] in Table 6.2-1.

**Table 6.2-1**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Payload size** | **Latency for model transmission (NOTE 1)** | **Transmission Data rate** |
| **LaneGCN** | 15 MByte | 3 seconds | 5 MByte/s |
| **ResNet-50** | 25 MByte | 3 seconds | 8.33 MByte/s |
| **ResNet-152** | 60 MByte | 3 seconds | 20 Mbyte/s |
| **PointPillar** | 18 MByte | 3 seconds | 6 MByte/s |
| **SECOND** | 20 MByte | 3 seconds | 6.67 MByte/s |
| **Part-A2-Free** | 226 MByte | 3 seconds | 75.33 MByte/s |
| **Part-A2-Anchor** | 244 MByte | 3 seconds | 81.33 MByte/s |
| **PV-RCNN** | 50 MByte | 3 seconds | 16.67 MByte/s |
| **Voxel R-CNN (Car)** | 28 MByte | 3 seconds | 9.33 MByte/s |
| **CaDDN (Mono)** | 774 MByte | 3 seconds | 248 MByte/s |
| NOTE 1: The transfer learning does not have a very high requirement for transmission latency since it is not a real-time inference service, hence it assumes the model transmission via direct device connection should be finished in 3 seconds. | | | |

# 7 Distributed/Federated Learning by leveraging direct device connection

## 7.1 Direct device connection assisted Federated Learning

### 7.1.1 Description

In many circumstances, an application server holding a Federated Learning (FL) task has a transmission delay requirement and limited FL coverage. An FL coverage means an area in which UEs the Application server can organize for federated learning.

An Application server has a transmission delay requirement for each FL member (UE). Some of UEs are actually holding valuable dataset but cannot fulfil transmission delay requirement, which leads to a decreasing of FL performance. However, if a UE’s direct network connection cannot fulfil the transmission delay requirement (i.e. an QoS on Uu), leveraging the devices with direct connections helps to involve more UEs holding valuable dataset for the FL task with the following case study:

A UE-A with bad transmission condition sends a UE’s training result to UE-B via direct device connection. In such case, a UE-B aggregates the training result locally and provides to UEs an update of training model for next round.

Some research e.g. in [6][7] have illustrated the increasing performance in subcase-B (we call it “decentralized averaging methods”). In order to include more devices to participate in FL and to reduce the devices’ reliance on the PS, the authors in [7] uses decentralized averaging methods to update the local ML model of each device. In particularly, using the decentralized averaging methods, each device only needs to transmit its local ML parameters to its neighboring devices. And the neighboring devices can use the acquired ML parameters to estimate the global ML model. Therefore, using the decentralized averaging methods can reduce the communication overhead of FL parameter transmission.

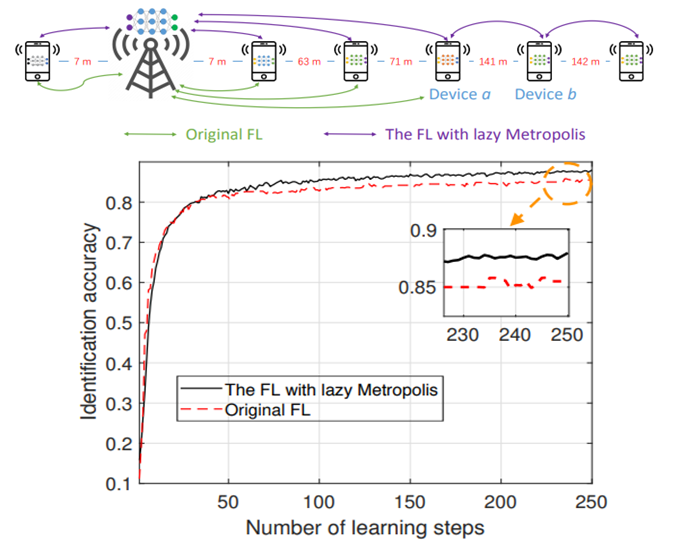


Figure 7.1-1 FL with decentralized averaging method outperforms the original FL

To show the performance of decentralized averaging method, the [6] implemented a preliminary simulation for a network that consists of one BS that is acted as an application server and six devices, as shown in Figure-1. In Figure-1, the green and purple lines respectively represent the local ML parameter transmission of original FL and the FL with decentralized averaging methods. Due to the transmission latency requirement, only 4 devices can participate in original FL. For the FL with the decentralized averaging update method, 6 devices can participate in the FL training process since the devices which are out of coverage can connect to their neighbouring devices (i.e. Device a and Device b) for model updating.

From Figure-1, we can see that the FL with decentralized averaging method outperforms the original FL in terms of identification accuracy. Specifically, the original FL (without using direct device connection) has an upper limit of identification accuracy to about 0.85, while using direct device connection for decentralized averaging method helps to increase the identification accuracy to about 0.88 which is actually a big optimization since the line already goes smoothly after 200 round of FL training.

Besides, the FL leveraging direct device connection can also reduce the energy consumption for some devices since it only needs to transmit its ML model parameters to device instead of the BS.

### 7.1.2 Pre-conditions



Figure 7.1-2 two UEs performs decentralized FL using Direct Device connection

As depicted in Figure-2, there is an Application server for federated learning which needs to communicate with the UEs in a FL coverage for FL task.

To achieve an ideal performance (i.e. fast convergence and high model accuracy), there is a transmission latency requirement to each FL member UE’s data transmission.

Alice and Bob are FL members but their cell phones sometimes have bad signal condition which cannot transmit data to FL service directly. Meanwhile, Bob is willing to support the “decentralized averaging method” (as described in clause 7.1.1) service for its neighbouring cell phones.

Alice, Bob are neighbouring to each other within a FL coverage.

### 7.1.3 Service Flows

1. Alice is a FL member and already acquires the global AI/ML model from the Application server for FL task, later on when Alice moves to a tunnel with bad signal condition, Alice cell phone’s with direct device connection with her neighboring cell phone cannot transmit model data to its Application server anymore.
2. In the tunnel, Alice discovers Bob, who is neighboring to Alice, a FL member and willing to activate the “decentralized averaging method” service. Thus, Alice requests Bob to establish a direct device connection so that Alice can transmit the AI/ML model training result to Bob.
3. Bob updates the AI/ML model based on Alice’s training result and Bob’s local training result. And Bob sends the updated AI/ML model to Alice for further training.

When Bob moves to a good coverage and is able to transmit the AIML training model (e.g. after several rounds of AIML model parameters exchange between Alice and Bob have been done), Bob transmits the training result to Application server to assist the Application server to perform a global model updating.

### 7.1.4 Post-conditions

By leveraging direct device connection, Alice and Bob can keep the model training of a FL task even when they are under a bad network coverage. And the training result between Alice and Bob can be further uploaded to Application server for global model updating.

Thanks to leveraging direct device connection, it helps FL to be performed even when no communication availability to FL server. Such use case helps to optimize the FL performance.

### 7.1.5 Existing features partly or fully covering the use case functionality

**In 3GPP TS 22.261 [8] v18.6.1 clause 6.40.2**

Based on operator policy, the 5G system shall be able to provide means to allow an authorized third-party to monitor the resource utilisation of the network service that is associated with the third-party.

NOTE 1: Resource utilization in the preceding requirement refers to measurements relevant to the UE’s performance such as the data throughput provided to the UE.

Based on operator policy, the 5G system shall be able to provide an indication about a planned change of bitrate, latency, or reliability for a QoS flow to an authorized 3rd party so that the 3rd party AI/ML application is able to adjust the application layer behaviour if time allows. The indication shall provide the anticipated time and location of the change, as well as the target QoS parameters.

Based on operator policy, 5G system shall be able to provide means to predict and expose predicted network condition changes (i.e. bitrate, latency, reliability) per UE, to an authorized third party.

Subject to user consent, operator policy and regulatory constraints, the 5G system shall be able to support a mechanism to expose monitoring and status information of an AI-ML session to a 3rd party AI/ML application.

NOTE 2: Such mechanism is needed for AI/ML application to determine an in-time transfer of AI/ML model.

Subject to user consent, operator policy and regulatory requirements, the 5G system shall be able to expose information (e.g. candidate UEs) to an authorized 3rd party to assist the 3rd party to determine member(s) of a group of UEs (e.g. UEs of a FL group).

### 7.1.6 Potential New Requirements needed to support the use case

#### 7.1.6.1 Functional requirement

[P.R.7.1-001] Based on user consent and operator policies, the 5G system shall be able to configure a group of UEs who participate in the same service group (e.g. for the same AI-ML FL task) to establish communication with each other via direct device connection e.g. when direct network connection cannot fulfil the required QoS.

[P.R.7.1-002] Based on user consent, operator policies and the request from an authorized 3rd party, the 5G system shall be able to dynamically add or remove UEs to/from the same service (e.g. a AI-ML federated learning task) when communicating via direct device connection.

#### 7.1.6.2 KPI requirement for direct device communication

The 5G system shall be able to support the following KPI for direct device connection as defined in Table 7.1.6-1

NOTE: The table refers to a typical AI/ML model for image recognition i.e. a 7-bit CNN model VGG16\_BN using 224×224×3 images as training data) [2].

Table 7.1.6-1: Latency and user experienced UL/DL data rates for uncompressed Federated Learning

| **Model size**  **(8 bit VGG-16 BN) (see NOTE 2)** | **Mini-batch size**  **(images)** | **Maximum latency for trained gradient uploading and global model distribution (see NOTE 1)** | **User experienced UL/DL data rate for trained gradient uploading and global model distribution (see NOTE 2)** |
| --- | --- | --- | --- |
| 132 Mbyte | 64 | 3.24s | 325Mbit/s |
| 32 | 1.9s | 55Mbit/s |
| 16 | 1.3s | 810Mbit/s |
| 8 | 1.1s | 960Mbit/s |
| 4 | 1.04s | 1.0Gbit/s |
| NOTE 1: Latency in this table is assumed 20 times the device GPU computation time for the given mini-batch size.  NOTE 2: Values provided in the table are calculative needs for an 8-bit VGG16 BN model with 132MByte size [2] | | | |

## 7.2 Asynchronous FL via direct device connection

### 7.2.1 Description

Federated Learning (FL) is an important machine learning service. Due to the Synchronous FL [8], Sync-FL, requires a strict communication quality for each UE in order to get all the intermediate results to FL server in time, the Synchronous FL is sometimes vulnerable to the unpredicted wireless condition and the divergence of UEs’ capabilities. Therefore, the Asynchronous FL [9], Async-FL, has been widely used in many circumstances. The main idea of Async-FL is to let UE report its result whenever it is ready and the FL server will refresh the model without waiting for all intermediates results are collected. The Sync-FL and Async-FL have pros and cons as the table 7.2-1 shows.

**Table 7.2-1 Comparison of Sync-FL and Async-FL**

|  |  |  |
| --- | --- | --- |
|  | **Sync-FL** | **Async-FL** |
| **Total computation workload** | Lower. | Higher. The UE will get a new model for training when it uploads the result without waiting for other UE’s result. So the computation work load in each UE can be increased. |
| **Communication requirement** | Higher. All UEs shall report its result before next FL round starts | Lower. |

### 7.2.2 Pre-conditions

The direct device connection can be used to realize the Async-FL. As The figure 7.2-1 shows, for some UEs who are in a bad coverage it can use the indirect network connection to communicate with Parameter Server (PS). The communication requirement via indirect network connection can be relaxed i.e. no need to transmitted all UEs training result with a restricted timing.

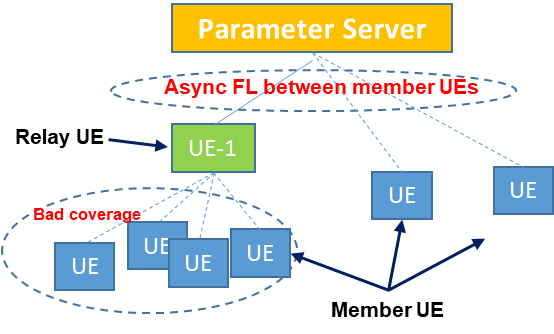


Figure 7.2-1 Group based Async-FL

For each member UE, it can send its training result to the PS via either direct network connection or indirect network connection, and the PS will send a new model to the member UE without waiting for other Aggregate UEs’ result (i.e. Async-FL).

### 7.2.3 Service Flows

1) The Parameter Server (PS) distributes the global model to FL member UEs via direct network connection or indirect network connection; For the UEs in bad coverage, it can use the indirect network connection to perform a Async-FL with PS.

2) When receiving the training result from member UE, the relay UE sends it to the Parameter server immediately to get a new model for the member UE. Due to the relay UE has a limited QoS for its own network connection (PDU session), the relay UE needs to determine the QoS for indirect network connection for each of member UEs based on an aggregated QoS (QoS upper limit) for the group of members served by the relay UE.

3) The Async-FL will be performed until the model accuracy reached a certain threshold.

### 7.2.4 Post-conditions

Thanks to the indirect network connection, the FL server can still use the valuable data stored in UEs who are out of coverage with the method of Async-FL. The model training is finally finished with expected model performance.

The charging for an Remote UE using an Indirect 3GPP Communication will be done.

### 7.2.5 Existing features partly or fully covering the use case functionality

**In TS 22.261 (v19.2.0) clause 3.1**

**aggregated QoS:** QoS requirement(s) that apply to the traffic of a group of UEs.

**In TS 22.261 (v19.2.0) clause 6.9**

The 5G system shall support different traffic flows of a remote UE to be relayed via different indirect network connection paths.

The connection between a remote UE and a relay UE shall be able to use 3GPP RAT or non-3GPP RAT and use licensed or unlicensed band.

The connection between a remote UE and a relay UE shall be able to use fixed broadband technology.

The 5G system shall be able to provide indication to a remote UE (alternatively, an authorized user) on the quality of currently available indirect network connection paths.

The 5G system shall be able to maintain service continuity of indirect network connection for a remote UE when the communication path to the network changes (i.e. change of one or more of the relay UEs, change of the gNB).

The 5G system shall be able to support a UE using simultaneous indirect and direct network connection mode.

The 5G system shall enable the network operator to authorize a UE to use indirect network connection. The authorization shall be able to be restricted to using only relay UEs belonging to the same network operator. The authorization shall be able to be restricted to only relay UEs belonging to the same application layer group.

**In TS 22.261 (v19.2.0) clause 6.40,**

Subject to user consent, operator policy and regulatory requirements, the 5G system shall be able to expose information (e.g. candidate UEs) to an authorized 3rd party to assist the 3rd party to determine member(s) of a group of UEs (e.g. UEs of a FL group).

**In TS 22.115 (V18.0.0) Clause 4.8 on "Charging Requirements for Indirect 3GPP Communication"**

This section describes the requirements to enable operator collection of charging data for an Evolved ProSe Remote UE and Relay UE using an Indirect 3GPP Communication. The requirements also apply in the roaming case.

The 3GPP core network shall be able to collect charging data for an Evolved ProSe Remote UE which accesses the 3GPP core network through an Indirect 3GPP Communication.

### 7.2.6 Potential New Requirements needed to support the use case

[P.R. 7.2-001] 5GS shall be able to support an aggregated QoS for a group UEs served by a relay UE.

[P.R. 7.2-002] 5GS shall be able to provision an aggregated QoS to a relay UE for a group-based service.

[P.E. 7.2-003] Based on 3rd party request and user consent, the 5G system shall be able to expose information (e.g. candidate UEs) to an authorized 3rd party to assist the 3rd party for UE member(s) selection in a group of UEs (e.g. UEs of a FL group), for UEs using direct or indirect network connection. E.g. the 3rd party request may include the expected QoS as a criteria for UE member selection.

## 7.3 5GS assisted distributed joint inference for 3D object detection

### 7.3.1 Description

Distributed joint inference is to leverage multiple nodes (e.g. UEs) to provide inference results so that an aggregation of those inference results can lead to a better performance.

When a 3rd party vehicle wants to obtain relevant information of a certain vehicle 1 (e.g. position, width, length, height, profile, orientation), the data collected by the 3rd party vehicle itself is limited. For example, as shown in figure 7.3.1-1, the 3rd party vehicle, which is directly behind vehicle 1, can only obtain relevant data on the tail of vehicle 1 through sensors, and can identify the width and height of the vehicle 1 through the inference of the local 3D object detection model, but there is no way to know the length of the vehicle 1, or even a more precise vehicle profile, orientation, etc. In addition, although the location of UE1 can be known through equipment such as the radar of the 3rd party vehicle, limited by the singleness of the data, the positioning accuracy based on the information obtained by a single vehicle is limited.



Figure-7.3.1-1: Joint inference among multiple vehicles for 3D object detection

All of the above problems need to be solved through multi-vehicle joint inference. The performance to use the joint inference is shown in the Figure-7.3.1-2. Its clear shows that despite the green vehicle generating false orientations and location by its local model, the global map (i.e., the red box) can correct the orientation and location error for the green vehicle based on the aggregated results of three vehicles (i.e. blue box, green box and yellow box) [23].

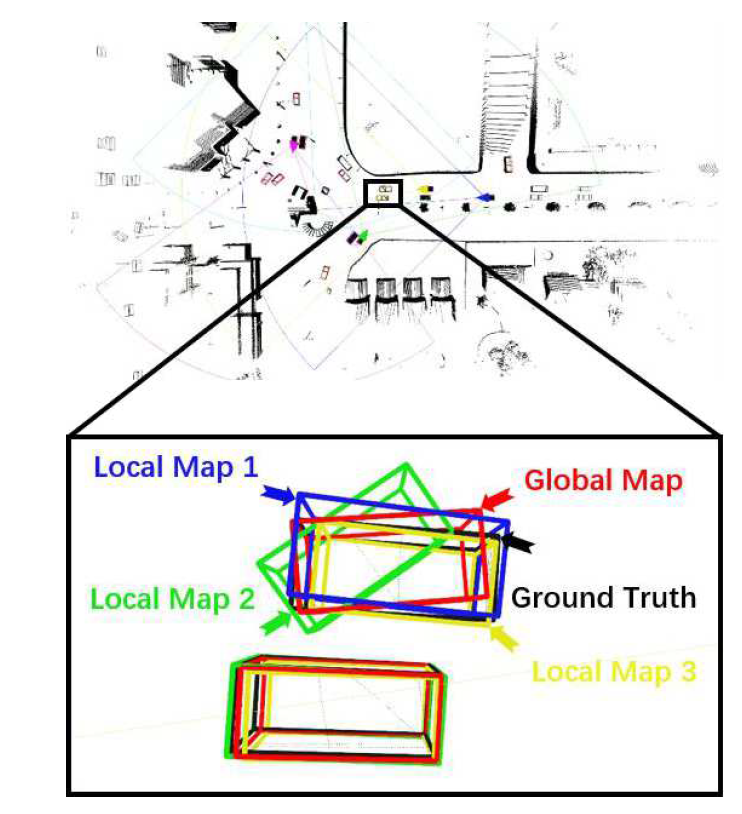


Figure-7.3.1-2: distributed joint learning leads to a better inference performance

### 7.3.2 Pre-conditions

As shown in figure-7.3.2-1, when a vehicle accident occurs somewhere and the road is congested. Alice's auto-driving vehicle wants to know the complete situation of the accident (i.e. the exact location and shape of the accident vehicle including the length, width and height of the vehicle), so as to use the inference result for auto-driving decision in real time. Alice’s vehicle needs to find and establish connections with vehicles located in different position to the accident vehicle, and collect inference results to perform the accurate 3D object detection of the accident vehicle.

Though the accident vehicle cannot move due to the collision to a barricade afront, the electronic device can still work as normal.



Figure-7.3.2-1: Joint inference among multiple vehicles for accident vehicle detection

### 7.3.3 Service Flows

1. Alice’s vehicle wants to know the complete situation of the accident vehicle. So, her car sends the request to 5G system to select the vehicles located in different positions /direction and a certain distance to the accident vehicle
2. Based on the candidate UE list located in different position to the accident vehicle, the Alice’s vehicle starts to establish direction device connection to each of the selected vehicles and transmit the 3D object detection model to them via direct device connection.
3. Alice’s vehicle receives the inference result the 3D object detection model made by the selected vehicles and further aggregate the results to acquire a highly accurate 3D object reconstruction of the accident vehicles.
4. Alice’s vehicle may also share the aggregated result to other vehicles or the application server so that they can use it to assist their own intelligent driving as well.

### 7.3.4 Post-conditions

Thanks to candidate UE list provided by the 5G system and inference results provided by other vehicles, Alice’s vehicle can get the situation of the accident scene accurately and make a path planning to avoid road congestion effectively.

### 7.3.5 Existing features partly or fully covering the use case functionality

In TS 22.261 clause 6.40.2, there is a requirement for FL scenario, i.e. the 5GS to assist 3rd party to determine FL members. But it is between the 5GS NF and the 3rd party. For the distributed joint inference use case, the communication is between the 3rd party UE and the UE1 or other UEs. The existing 5G system cannot support to help find the suitable UEs request by the 3rd party UE via direct device connection.

*Subject to user consent, operator policy and regulatory requirements, the 5G system shall be able to expose information (e.g. candidate UEs) to an authorized 3rd party to assist the 3rd party to determine member(s) of a group of UEs (e.g. UEs of a FL group).*

### 7.3.6 Potential New Requirements needed to support the use case

#### 7.3.6.1 Potential Functionality Requirements

[P.R.7.3-001] Subject to user consent, operator policy, and 3rd party’s request, 5GS shall be able to provide and configure the QoS applied to a group of UEs communicating via direct device connection (e.g. part of a joint AIML inference task).

NOTE: the above requirement assumes unicast type of communication.

[P.R.7.3-002] Subject to user consent, operator policy and 3rd party’s request, the 5G system shall be able to provide information of certain UEs (e.g. located in a specific location) to an authorized 3rd party (e.g. to assist a joint AIML task using direct device communication).

#### 7.3.6.2 Potential KPI Requirements

According to [24], some typical 3D objection model size and transmission KPI are listed in the table below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Type** | **Max allowed DL end-to-end latency** | **Experienced data rate**  **(PC5)** | **Model size** | **Communication service availability** |
| PointPillar | 1s | 0.14 Gbit/s | 18 MByte | 99.99 % |
| SECOND | 1s | 0.16 Gbit/s | 20 MByte | 99.99 % |
| PV-RCNN | 1s | 0.4 Gbit/s | 50 MByte | 99.99 % |
| Voxel R-CNN (Car) | 1s | 0.22 Gbit/s | 28 MByte | 99.99 % |

# 8 Consolidated potential requirements and KPIs

## 8.1 Consolidated potential requirements

### 8.1.1 Enabling

Table 8.1.1 – Enabling Consolidated Requirements

| CPR # | Consolidated Potential Requirement | Original PR # | Comment |
| --- | --- | --- | --- |
| **CPR 8.1.1-1** | Based on user consent, operator policy and 3rd party request, the 5G system shall support a means to authorize specific UEs to transmit data (e.g. AI/ML model intermediate data for a specific application, AI/ML model transfer for a specific application) via direct divice connection in a certain location and time. | P.R.5.2.6-001  P.R.6.2-001 |  |
| **CPR 8.1.1-2** | Based on user consent, operator policy, and 3rd party’s request, the 5G system shall be able to provide means for a operator to authorize a group of UEs who participate in the same service group (e.g. for the same AI-ML FL task) to exchange data with each other via direct device connection e.g. when direct network connection cannot fulfil the required QoS. | P.R.6.1-002  P.R.7.1-001 |  |
| **CPR 8.1.2-3** | Based on user consent, operator policies and the request from an authorized 3rd party, the 5G network shall be able to dynamically add or remove UEs to/from the same service (e.g. a AI-ML federated learning task) when communicating via direct device connection. | P.R.7.1-002 |  |

### 8.1.2 QoS support

Table 8.1.2 – QoS support Consolidated Requirements

| CPR # | Consolidated Potential Requirement | Original PR # | Comment |
| --- | --- | --- | --- |
| **CPR 8.1.2-1** | Based on user consent and operator policy, the 5G network shall be able to have the means for operator to modify the communication QoS so that the end to end latency can be satisfied when a relay UE is involved for a proximity-based work task offloading.  NOTE 1: Due to the proximity-based work task offloading, the data size transmitted via sidelink and Uu of the indirect network connection can be different  NOTE 2: The split point selection is dynamic. In consequence, the amount of intermediate data will vary.  NOTE 3: Similarly as existing direct device connection based transmission the QoS for the specific service can be impacted by the distance between UEs then the UE reselection may be needed. | P.R.5.1.6-001  P.R.5.2.6-005 |  |
| **CPR 8.1.2-2** | Subject to user consent and operator policy, the 5G network shall be able to configure the QoS (latency, reliability, data rate) of the communication path of the direct device connection. | P.R.5.2.6-004 |  |
| **CPR 8.1.2-3** | Based on user consent, operator policy and regulatory requirement, the 5G network shall be able to support means to monitor the characteristics (e.g. experienced data rate, latency) of the traffic transmitted via direct device connection or the traffic relayed by a UE and expose the monitoring information to trusted 3rd party.  NOTE: The monitoring information doesn’t include any user position-related data. | P.R.6.1-001 |  |
| **CPR 8.1.2-4** | Subject to user consent, operator policy and trusted 3rd party request, the 5G system shall be able to provide means for operator to predict and expose condition changes of QoS for indirect network connection (i.e. bitrate, latency, reliability). | P.R.5.2.6-002 |  |
| **CPR 8.1.2-5** | The 5G system shall support a mechanism for an authorized third-party to negotiate with a 5G network for a suitable QoS for direct device connections for a group of UEs to exchange data with each other | P.R.6.1-003 |  |
| **CPR 8.1.2-6** | Based on user consent, operator policy and trusted 3rd party’s request, the 5G network shall be able to support and provision an aggregated QoS for a group of UEs served by a relay UE. | P.R. 7.2-001  P.R. 7.2-002 |  |
| **CPR 8.1.2-7** | Based on user consent, operator policy and 3rd party’s request, the 5G network shall be able to provide and configure the QoS limitation applied to a group of UEs communicating via direct device connection (e.g. part of a joint AIML inference task).  NOTE: the above requirement assumes unicast type of communication. | P.R.7.3-001 |  |

### 8.1.3 Member selection

Table 8.1.3 – Member selection Consolidated Requirements

| CPR # | Consolidated Potential Requirement | Original PR # | Comment |
| --- | --- | --- | --- |
| **CPR 8.1.3-1** | Subject to user consent and operator policy, the 5G network shall be able to expose information to a trusted 3rd party to assist the 3rd party to determine candidate UEs for data transmission via direct device connection (e.g. for AIML model transfer for a specific application).  NOTE: the information does not include user’s specific positioning and can be QoS relevant information | P.R.6.2-002 |  |
| **CPR 8.1.3-2** | Subject to user consent, operator policy, regulation and 3rd party’s request, the 5G network shall be able to provide information of certain UEs joining a certain AIML service (e.g. located in a specific location area) to a trusted 3rd party (e.g. to assist a joint AIML task using direct device communication)  NOTE: the information does not include user’s exact positioning information. | P.R.7.3-002 |  |

### 8.1.4 Charging

Table 8.1.4 – Charging Consolidated Requirements

| CPR # | Consolidated Potential Requirement | Original PR # | Comment |
| --- | --- | --- | --- |
| **CPR 8.1.4-1** | The 5G system shall be able to support charging mechanism for multiple UE exchange data for the same service using the direct device connection. | PR.5.1.6-003 |  |

## 8.2 Consolidated potential KPIs

### 8.2.1 Split AI/ML operation between AI/ML endpoints

Table 8.1-1 KPI Table of Split AI/ML operation between AI/ML endpoints for AI inference by leveraging direct device connection

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Max allowed UL end-to-end latency (NOTE 1)** | **UL Payload size (Intermediate data size)** | **UL Experienced data rate** | **Service area dimension** | **Communication service availability** | **Reliability** | **Remarks** |
| 2–100 ms | 0.1-1.5 Mbyte for each frame | 4.8-720 Mbps |  |  |  | Proximity-based work task offloading for Remote driving, AR displaying/gaming, remote-controlled robotics, video recognition and One-shot object recognition |
| 10 ms | 0.000016 – 1.6 MByte  (8 bits data format) | 0.128 - 1.28 Gbps | 900 m2 (30 m x 30 m) | 99.999 % | 99.999 % | Local AI/ML model split on factory robots |
| 10 ms | 0.000064 – 6.4 Mbyte  (32 bits data format) | 0.512 - 5.12 Gbps | Local AI/ML model split on factory robots |
| NOTE 1: For AIML model splitting, the UL intermediate data size is much bigger than DL result data size. | | | | | | |

### 8.2.2 AI/ML model/data distribution and sharing by leveraging direct device connection

Table 8.1-1 KPI Table of AI/ML model/data distribution and sharing by leveraging direct device connection

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Max allowed DL end-to-end latency**  **(NOTE 1)** | **Experienced DL data rate** | **DL Payload size** | **Communication service availability** | **Remark** |
| 1s | 0.14 - 1.92 Gbit/s | 18 - 240 MByte | 99.9 % | AI Model Transfer Management through Direct Device Connection |
| 3s | 5 – 81.33 Mbyte/s | 15 – 244 MByte | - | transfer learning for trajectory prediction |
| NOTE 1: the model distribution mainly requires an KPI for DL data transmission  NOTE 2: The AI/ML model data distribution is for a specific application service | | | | |

### 8.2.3 Distributed/Federated Learning by leveraging direct device connection

Table 8.1-1 KPI Table of Distributed/Federated Learning by leveraging direct device connection

| **Ul/DL Payload size**  **(NOTE 1)** | **Maximum UL/DL latency** | **Experienced UL/DL data rate** | **Reliability** | **Remark** |
| --- | --- | --- | --- | --- |
| 132 MByte | 2-3 s | 352 – 528 Mbit/s |  | Direct device connection assisted Federated Learning (Uncompressed model) |
| 132 MByte |  |  |  | Asynchronous Federated Learning via direct device connection |
| 18 – 50 MByte | 1 s | 140 – 220 Mbit/s | 99.99% |  |
| NOTE 1: For Distributed/federated learning, the UL and DL data size are almost the same due to the training gradient (UL) and global model (DL) are corresponding to the same model and have the same number of parameters | | | | |

# 9 Conclusion and recommendations

Regarding the Feasibility Study on traffic characteristics and performance requirements for AI/ML Model Transfer via direct device connection, the TR analyses use cases of AIML-Ph2 as follows:

* Use cases on split AI/ML operation between AI/ML endpoints for AI inference by leveraging direct device connection:
  + Proximity based work task offloading for AI/ML inference
  + Local AI/ML model split on factory robots.
* Use cases on AI/ML model/data distribution and sharing by leveraging direct device connection:
  + AI Model Transfer Management through Direct Device Connection;
  + 5GS assisted transfer learning for trajectory prediction.
* Use cases on Distributed/Federated Learning by leveraging direct device connection:
  + Direct device connection assisted Federated Learning;
  + Asynchronous FL via direct device connection;
  + 5GS assisted distributed joint inference for 3D object detection;

It is recommended to proceed with normative work including the potential new requirements identified by this TR. The consolidated potential requirements in Clause 8 are the baseline for the subsequent normative work.

Annex A:  
Change history

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Change history** | | | | | | | |
| **Date** | **Meeting** | **TDoc** | **CR** | **Rev** | **Cat** | **Subject/Comment** | **New version** |
| 2022.05 | SA1#98e | S1-221010 | - | - | - | Initial Skeleton | 0.0.0 |
| 2022.09 | SA1#99e |  | - | - | - | Inclusion of pCRs agreed at SA1#99: S1-222397; S1-222158; S1-222398; S1-222399; S1-222400; S1-222401 | 0.1.0 |
| 2022.11 | SA1#100 |  |  |  |  | Inclusion of pCRs agreed at SA1#100: S1-223629; S1-223630; S1-223713; S1-223732 | 0.2.0 |
| 2023.02 | SA1#101 |  |  |  |  | Inclusion of pCRs agreed at SA1#101: S1-230783; S1-230393; S1-230394; S1-230087; S1-230396; S1-230784; S1-230744 | 0.3.0 |
| 2023-03 | SA#99 | SP-230224 |  |  |  | MCC clean-up for presentation to SA#99 | 1.0.0 |
| 2023-05 | SA1#102 |  |  |  |  | Inclusion of pCR agreed at SA1#102： S1-231104, S1-231509, S1-231797, S1-231741, S1-231512 | 1.1.0 |
| 2023-06 | SA#100 | SP-230513 |  |  |  | MCC clean-up for approval by SA#10 | 2.0.0 |
| 2023-06 | SA#100 | SP-230513 |  |  |  | Raised to v.19.0.0 by MCC following approval by SA#10 | 19.0.0 |