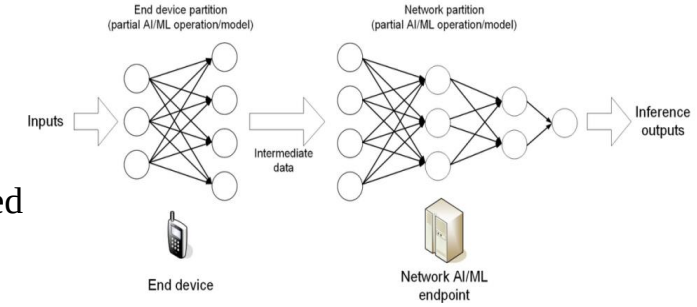
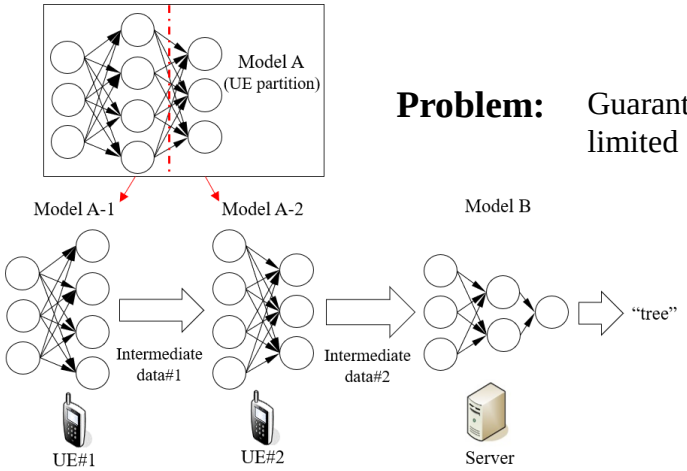
D2D-aided model splitting

Scenario:

AI/ML operation splitting between AI/ML endpoints

Network partition is fixed



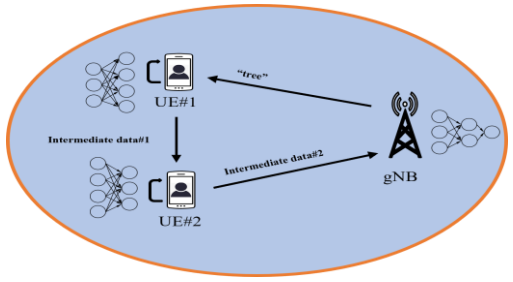


Problem: Guarantee the service when UE with limited computational capabilities

Solution: Offload part of Model A from UE1 to UE2

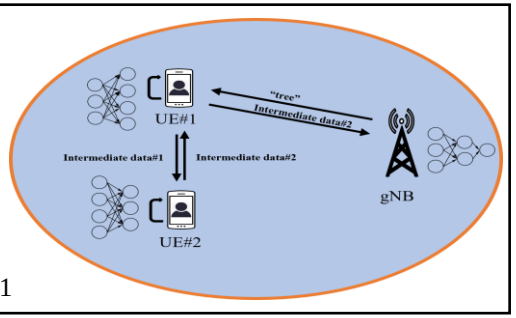
Scenario 1:
1). Good sidelink (UE 1 <--> UE 2)
2). Good up/down-link (UE 1 <--> gNB)
3). Good up/down-link (UE 2 <--> gNB)
4). Limited computation at UE#1
5). Sufficient computation at UE#2

Service flow:
UE#1 → UE#2 → gNB → UE#1



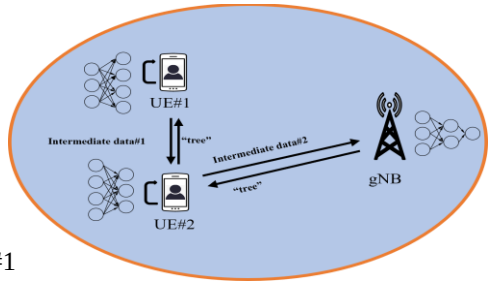
Scenario 2:
1). Good sidelink (UE 1 <--> UE 2)
2). Good up/down-link (UE 1 <--> gNB)
3). Bad up/down-link (UE 2 <--> gNB)
4). Limited computation at UE#1
5). Sufficient computation at UE#2

Service flow:
UE#1 → UE#2 → UE#1 → gNB → UE#1



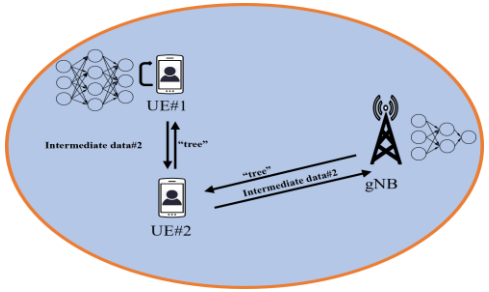
Scenario 3:
1). Good sidelink (UE 1 <--> UE 2)
2). Bad up/down-link (UE 1 <--> gNB)
3). Good up/down-link (UE 2 <--> gNB)
4). Limited computation at UE#1
5). Sufficient computation at UE#2

Service flow:
UE#1 → UE#2 → gNB → UE#2 → UE#1



Scenario 4:
1). Good sidelink (UE 1 <--> UE 2)
2). Bad up/down-link (UE 1 <--> gNB)
3). Good up/down-link (UE 2 <--> gNB)
4). Sufficient computation at UE#1
5). Sufficient or limited computation at UE#2

Service flow:
UE#1 → UE#2 → gNB → UE#2 → UE#1



Standard impact: **S1-220183** Study on AI/ML Model Transfer_Phase2

Objective: **Distributed AI training/inference based on Device to Device connection**, e.g. traffic KPIs, different QoS and functional requirements on slidelink transmission.