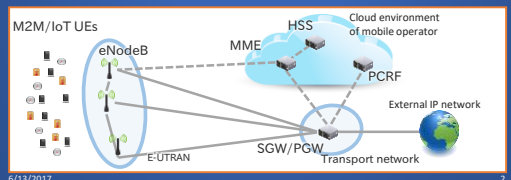
Design and performance evaluation of bearer aggregation method in mobile core network with C/U plane separation

○Shuya Abe, Go Hasegawa, Masayuki Murata
Osaka University

Background 1/2

- Cellular networks have to accommodate massive M2M/IoT communications
- Accommodating massive M2M/IoT communication increases load on mobile core network nodes
  - Signaling procedure to establish per-UE bearers which are logical paths for data packets before UEs start data communication
  - For bearer establishment, various signaling messages propagate between the nodes and are processed on the nodes

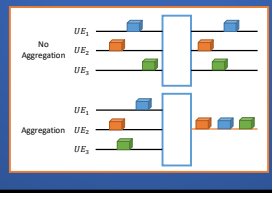


Background 2/2

- Various methods for accommodation of massive M2M/IoT terminals
  - server virtualization, C/U plane separation, SDN, etc.
  - Signaling procedure and network slices dedicated to M2M/IoT communication
  - Conveying data in C-plane signaling messages
- These existing methods
  - Can utilize server and network resources more efficiently
  - Can not reduce overhead required for bearer establishment itself
- Existing researches on bearer establishment overhead are based on the number and size of the signaling messages
  - Actually, the load of signaling messages are determined by many factors

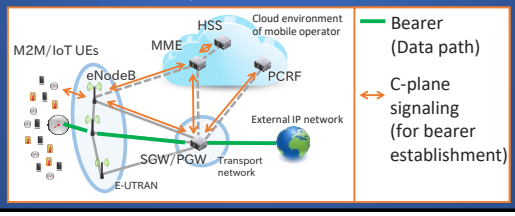
Purpose

- Performance evaluation of the mobile core network considering signaling processing load
  - Exploits the number of statements for processing signaling messages
- Bearer aggregation method is proposed and evaluated
  - Shares a bearer by multiple UEs
  - Decreases the load of bearer establishment



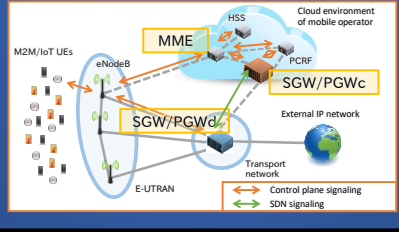
Network Model

UE	: Terminal accommodated to the mobile core network
eNodeB	: Base station for LTE
MME	: Pillar of signaling procedure for bearer establishment, Mobility management, Authentication, etc.
SGW/PGW	: Anchor point for intra-LTE mobility/ Gateway to external IP networks



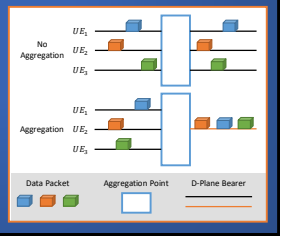
C/U Plane Separation

- Separating SGW/PGW node into **C-plane** and **U-plane** nodes and moving **C-plane** node to the cloud
  - C/U plane separation decreases propagation delay of signaling messages between cloud nodes
  - Server resources can be shared among cloud nodes
  - Additional signaling messages between SGW/PGWc and SGW/PGWd



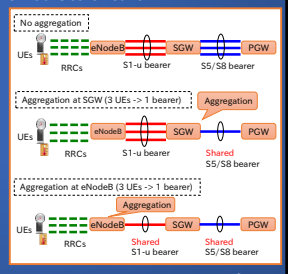
Bearer Aggregation

- Sharing a bearer by multiple UEs
  - Decreases signaling messages and node processing for establishing per-UE bearers
  - Increases signaling messages and processing for bearer aggregation
- Aggregation Point
  - SGW
  - eNodeB
- Aggregation Timing
  - When UEs attach (connect) to the mobile core network
  - When UEs start communication



Aggregation Point

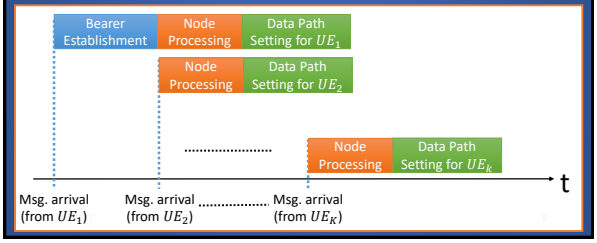
- Aggregation at SGW
  - Shares bearers between SGW and PGW (S5/S8 bearer)
  - Requires small modification to the mobile core network
    - S5/S8 bearer is kept established when UE becomes idle
- Aggregation at eNodeB
  - Shares bearers between eNodeB and SGW (S1-u bearer)
  - S5/S8 bearers are also aggregated
  - Requires large modification to the mobile core network
    - Shared S1-u bearer must be kept established until all UEs become idle



Aggregation Timing 1/2

Pre-determined Aggregation

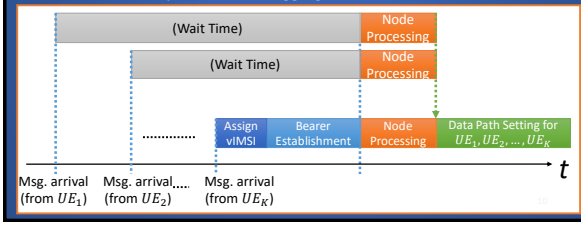
- A UE is associated with a shared bearer when the UE attaches to the network
- A shared bearer is activated when it is not established
- The data path setting to data plane nodes is necessary for each UE



Aggregation Timing 2/2

On-demand Aggregation

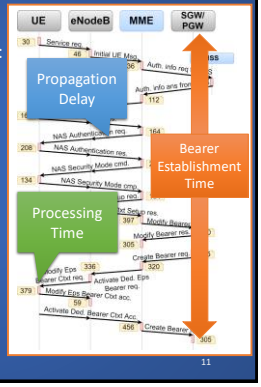
- A UE is associated with a shared bearer when the UE starts data communication
- Each UE must wait for the communication requests until the number of requests reaches the number of UEs in each group (aggregation level)
- Only one data path setting procedure is necessary when the number of requests reaches aggregation level



Performance Analysis

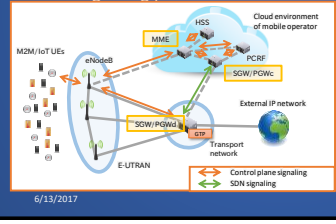
- Evaluation metrics
  - Bearer establishment time (T): Execution time required for the signaling procedure
- T is the sum of the following times
  - T_τ: Propagation delay
  - T_t: Processing time
  - T_w: Waiting time when on-demand aggregation

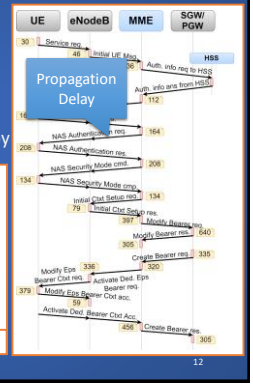
$$T = T_\tau + T_t + T_w$$



Propagation Delay

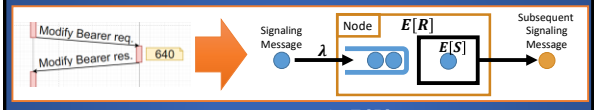
- Propagation delay depends on distance between nodes
  - In the cloud environment: small delay
  - Between cloud nodes and E-UTRAN / Transport network: large delay
- The number of propagations of signaling messages is determined by the signaling procedure





Processing Time

- M/G/1/PS queuing model is exploited



- Average processing time at a node $E[R]$

$$E[R] = \frac{\rho^r}{1-\rho}\frac{E[S^2]}{2E[S]} + \frac{1-\rho^r}{1-\rho}E[S]$$

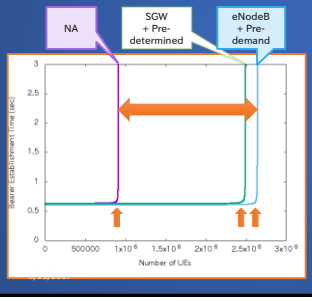
Variable in queuing model	Related Parameter in evaluation
λ: Arrival rate (of signaling messages)	# of UEs, # of processing, # of the nodes, UE's communication frequency
E[S]: Mean workload (Ave. signaling processing time)	Ave. # of statements, Server resources
ρ: System utilization	λ, E[S]

Evaluation Results

- Aggregation methods for evaluation combine aggregation point and aggregation timing
- Aggregation level (# of UEs for each shared bearer): 64
- Server resources and propagation delay are determined assuming a nation-wide mobile core network

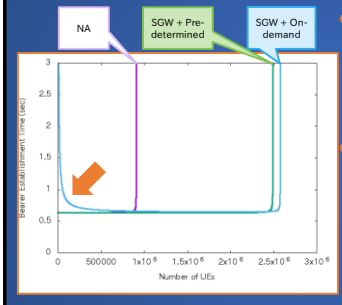
Effect of Aggregation Point

- Bearer establishment time increases sharply with a certain number of accommodated UEs
  - Defined as ``network capacity''



- Network capacity further increases by applying a bearer aggregation method
- Aggregation at eNodeB has larger network capacity than that at SGW
  - The number of total bearers in the network is smaller

Effect of Aggregation Timing



- On-demand aggregation (OA) has larger network capacity than Pre-determined
  - With OA, only one data path setting procedure is necessary for each group of UEs
- OA has a large bearer establishment time when the number of UEs is small
  - Caused by waiting time

Discussion

- Recommended combinations of aggregation point and timing is determined depending on the number of UEs and the mobility of UEs
- The amount of modification depends on the aggregation method

UEs' characteristics	Point	Timing	Required modification	Network capacity
high mobility	SGW	Pre-determined	small	low
massive, high mobility	SGW	On-demand	small	medium
low/no mobility	eNodeB	Pre-determined	large	high

Conclusion

- Summary
  - Evaluation of bearer aggregation methods to aggregate multiple M2M/IoT communication
  - Bearer aggregation methods improve the performance of the mobile core network
  - Appropriate aggregation method is determined depending on the characteristics of accommodated UEs
- Future work
  - C/U plane separation at eNodeB
  - Evaluation of U-plane performance
  - Experimental evaluation to support the results

Backup Slide

Propagation Delay

Nodes	Delay
UE – eNodeB	20 msec
eNodeB – SGW/PGWd	7.5 msec
eNodeB – MME	10 msec
SGW/PGWd – SGW/PGWc	10 msec
MME – SGW/PGWc	1 msec
GTP module – SGW/PGWd	1 msec

Server Resources

Node	Server Resources
UE	3,000 statements/sec
eNodeB	1,500 statements/sec
MME	3,000,000 statements/sec
SGW/PGWc	3,000,000 statements/sec
SGW/PGWd	3,000,000 statements/sec
GTP module	600,000 statements/sec