NOF 2017 London

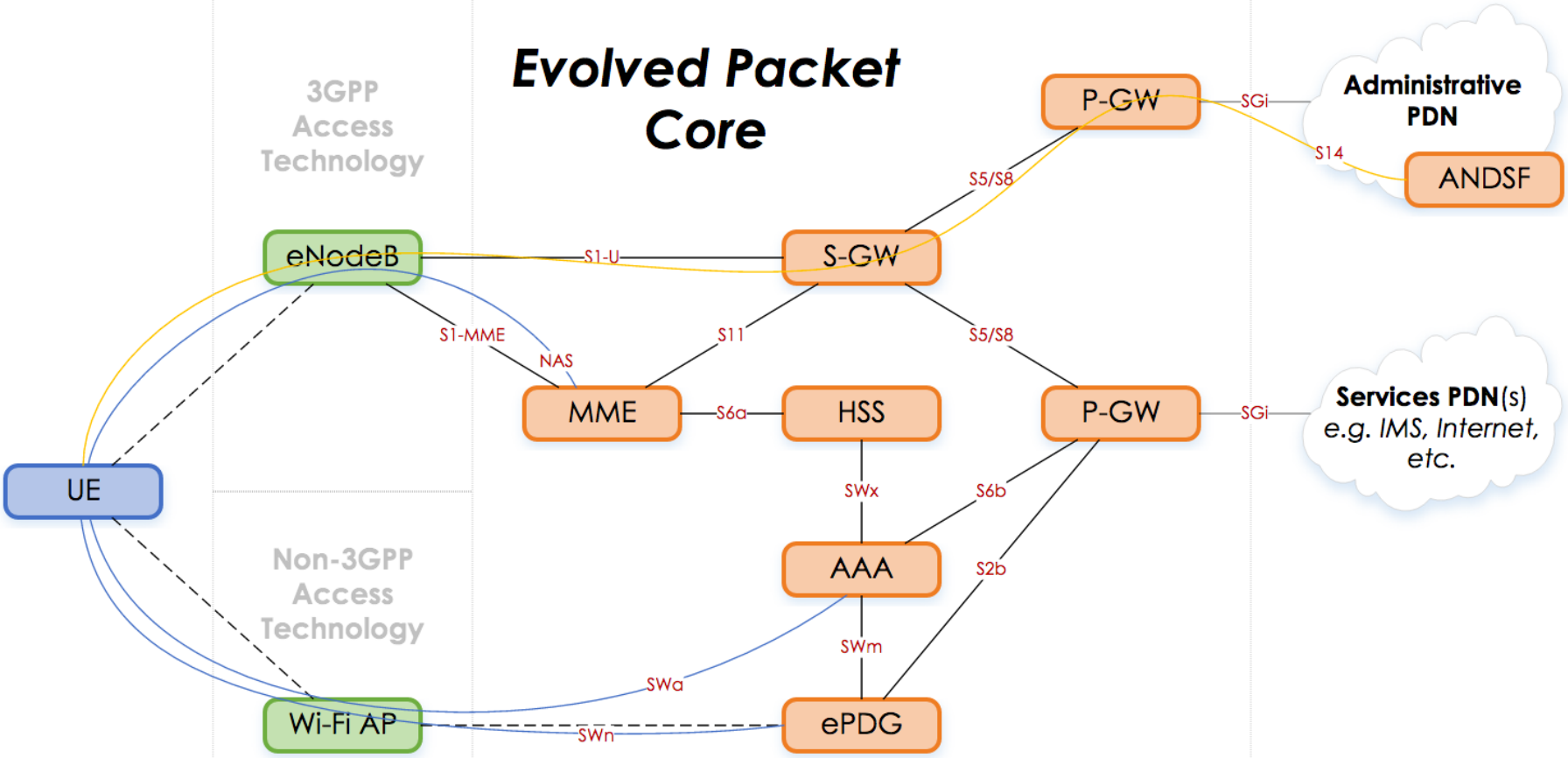
VIRTUALIZATION IN 5G SYSTEMS

PART II

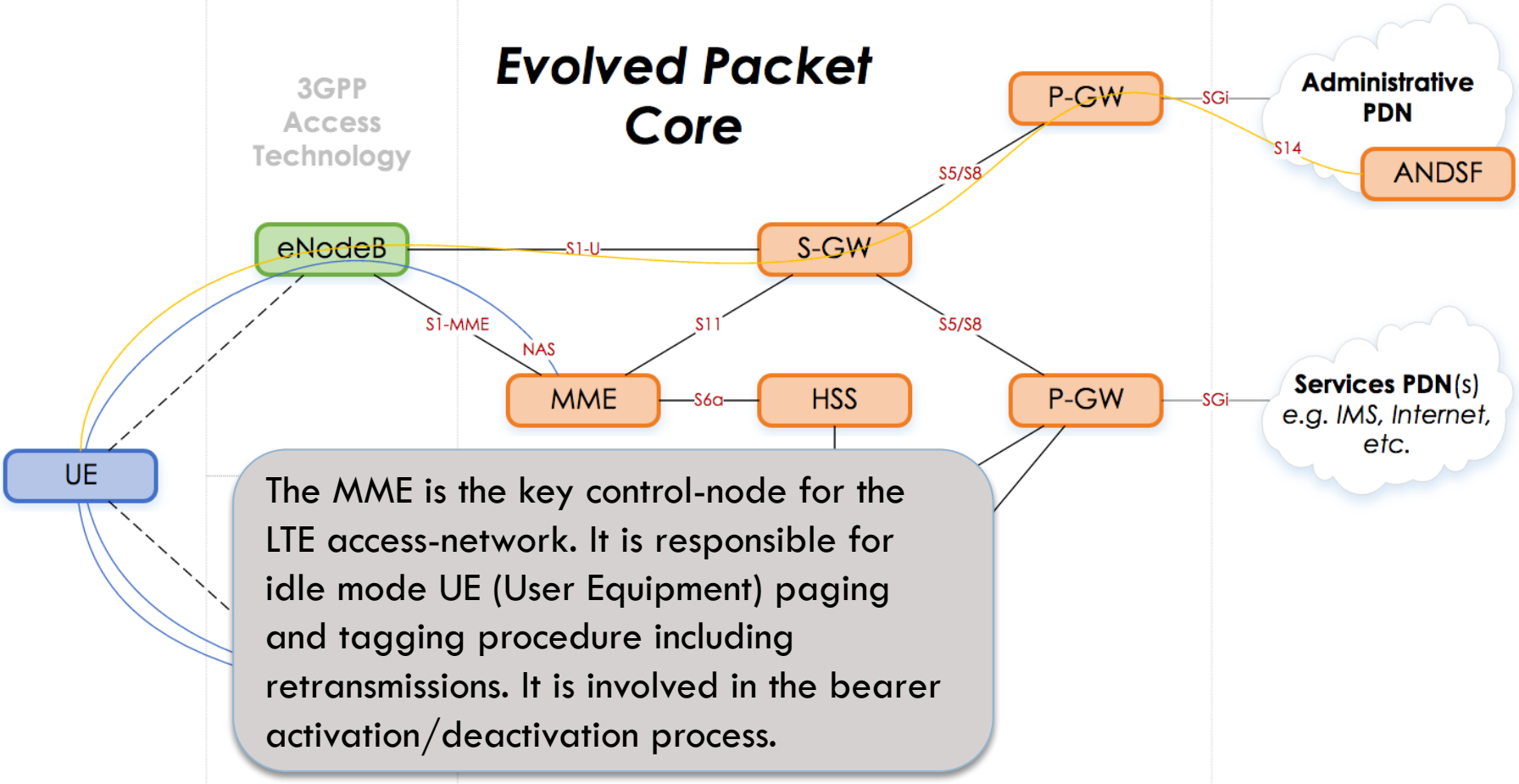
Fabrizio Granelli fabrizio.granelli@unitn.it

What do we have right now? 4G

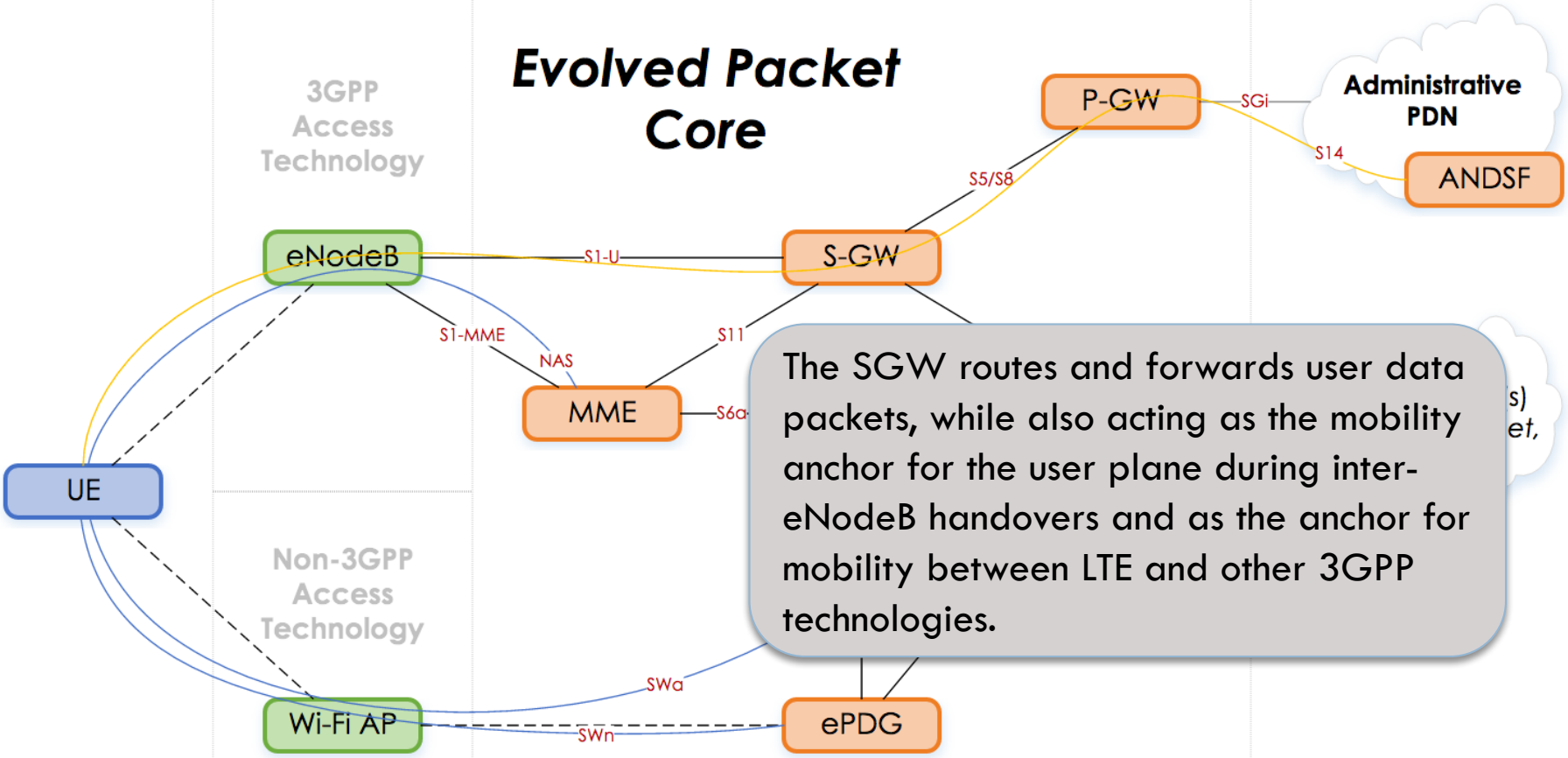
LTE in a nutshell



LTE in a nutshell



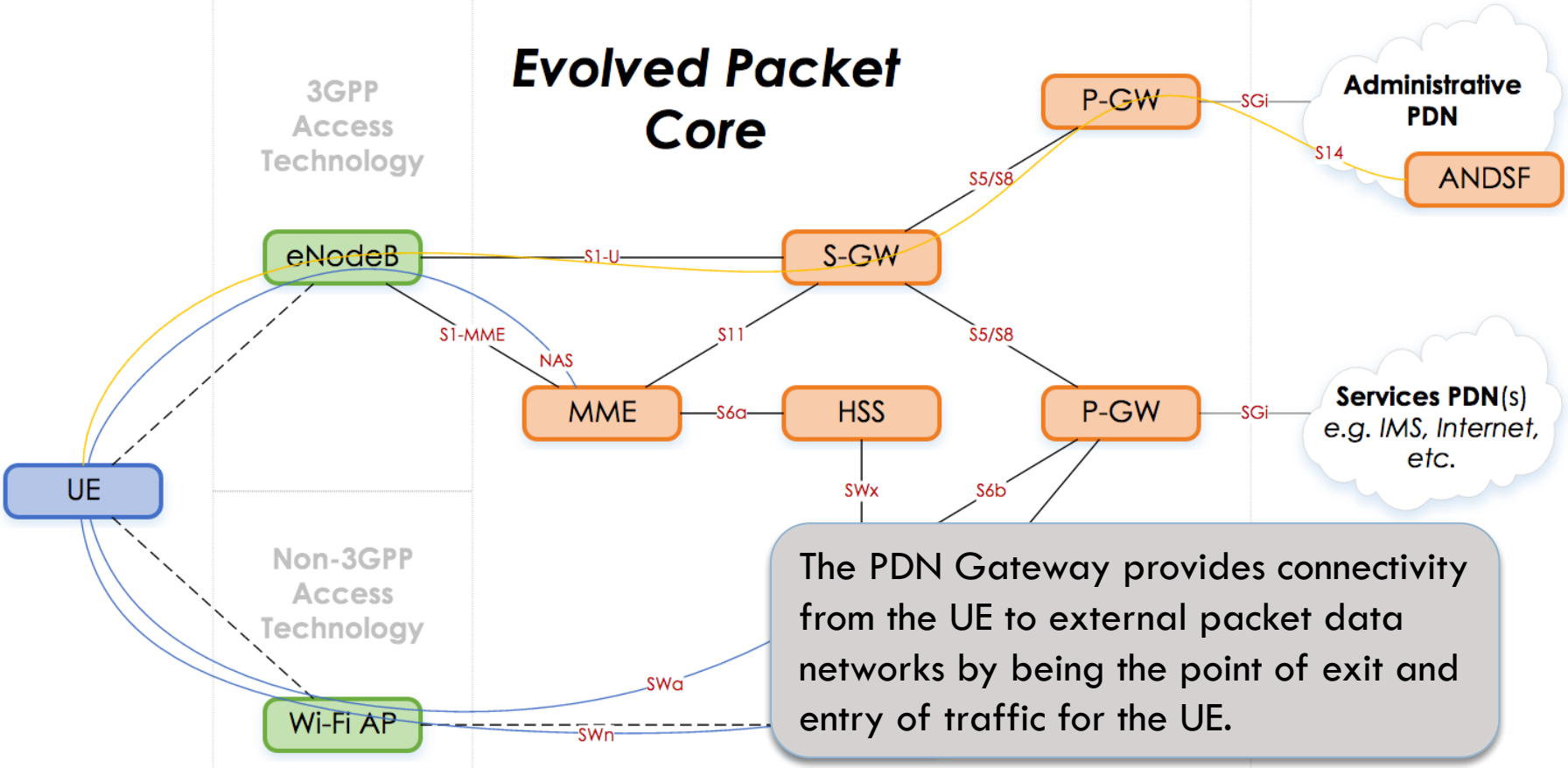
LTE in a nutshell



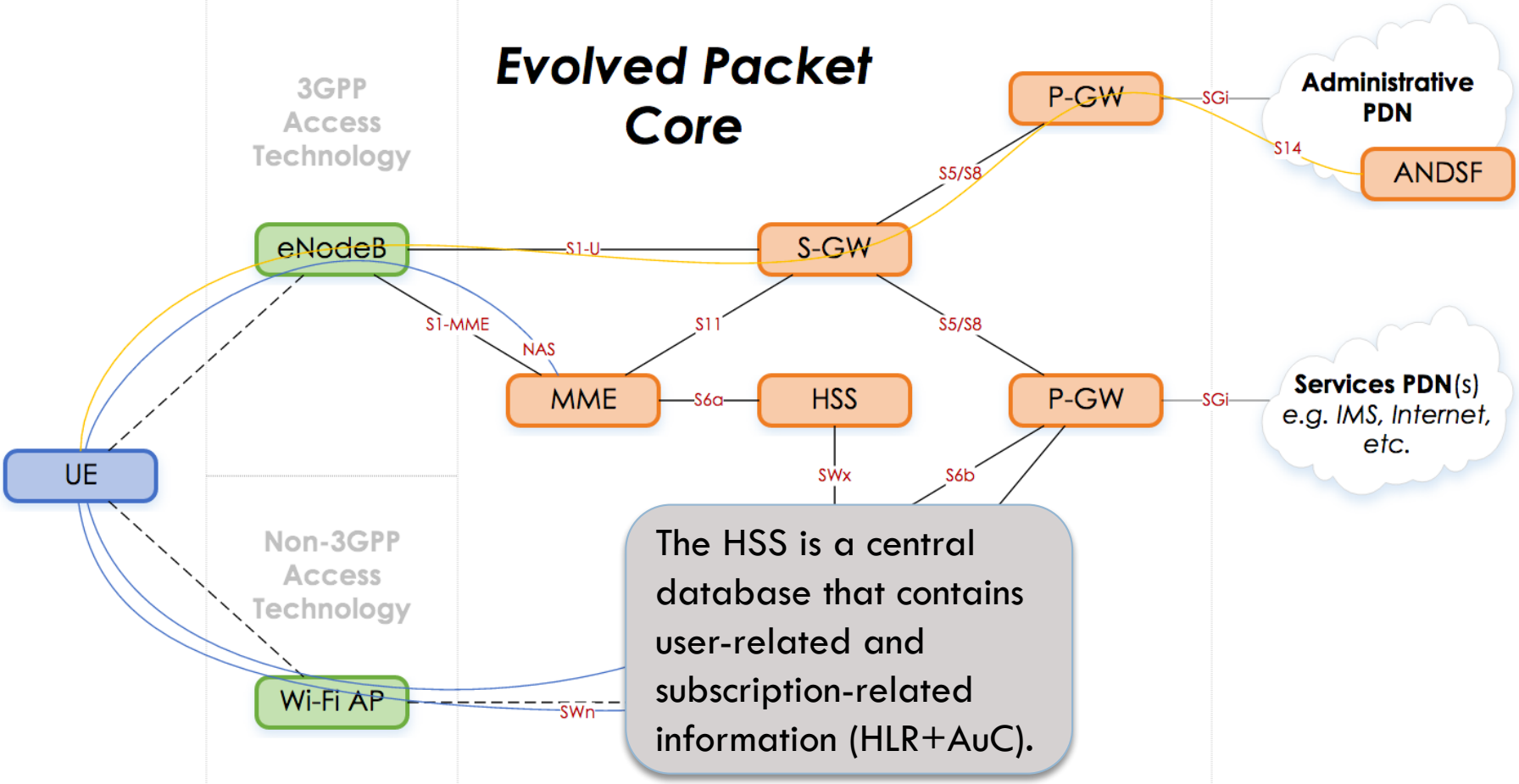
By Joe Deu-Ngoc - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=37142420

NOF 2017 London

LTE in a nutshell



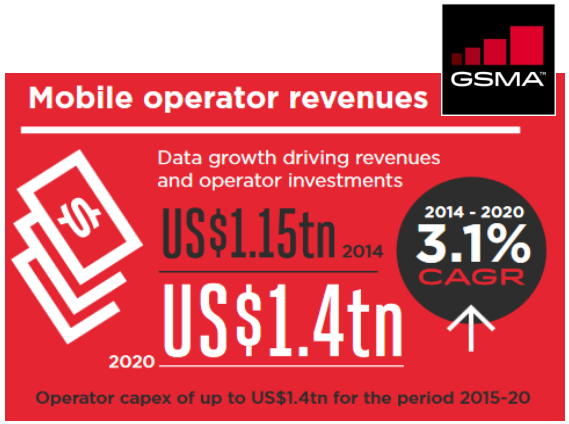
LTE in a nutshell



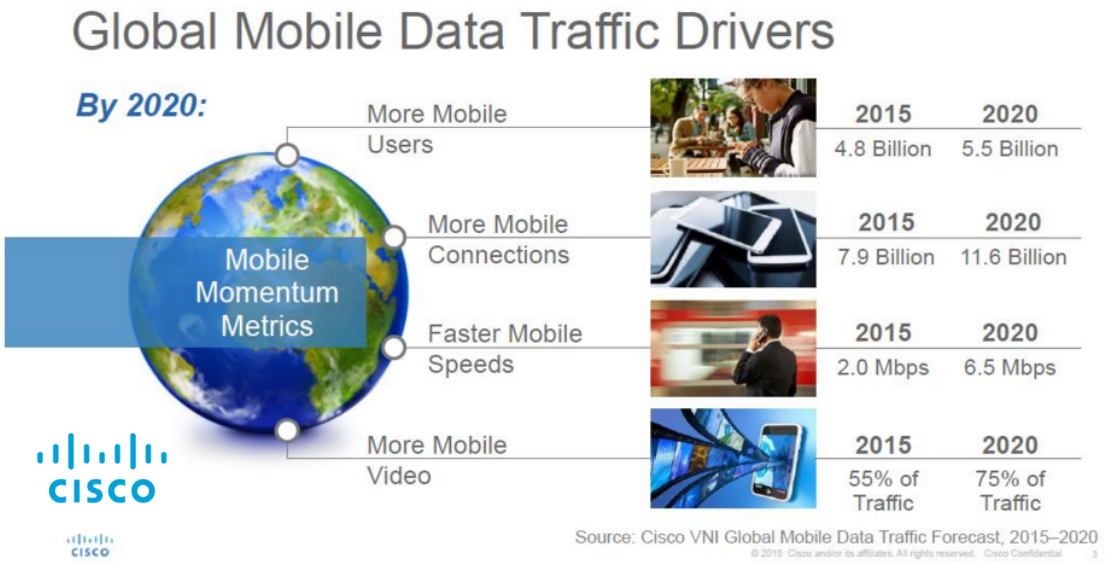
Why virtualization matters in 5G Era?

Mobile Traffic and Business Trends 2020

- Increased traffic levels / Increased number of devices
  - Global mobile data traffic is expected to reach 30.6 exabytes monthly by 2020, an 8-fold increase as compared to 2015
  - 5G networks will connect over 1billion devices by 2020



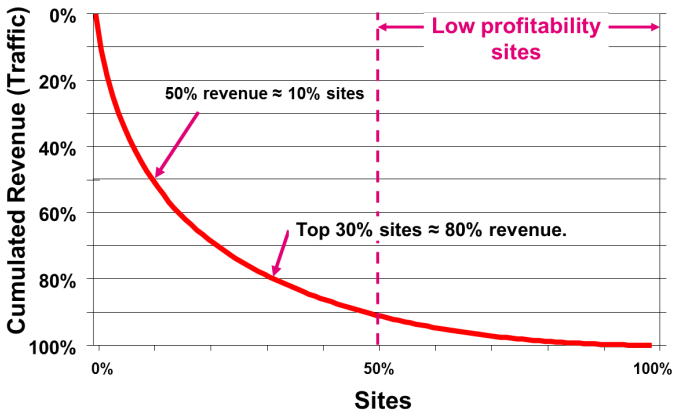
CAGR - Compound Annual Growth Rate

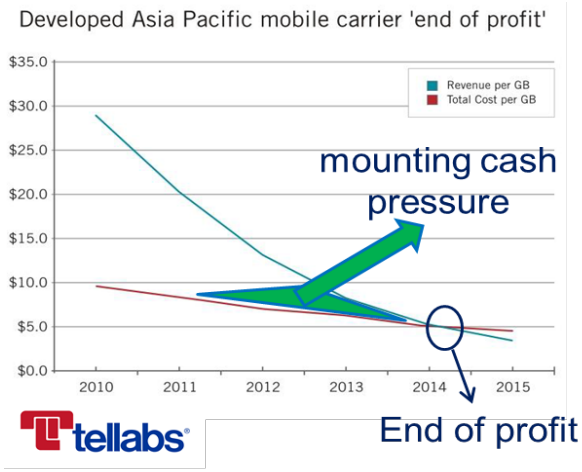


NOF 2017 London

The need for Network Sharing

- Share the underutilized resources
  - 50% of revenue created by < 10% of sites
  - diverse traffic patterns – mobility – cause resource underutilization
- RAN Sharing can recover significant OPEX/CAPEX costs
  - up to 20% of OPEX for typical European operator
  - reduce CAPEX in developing countries, e.g. up to 70% in India

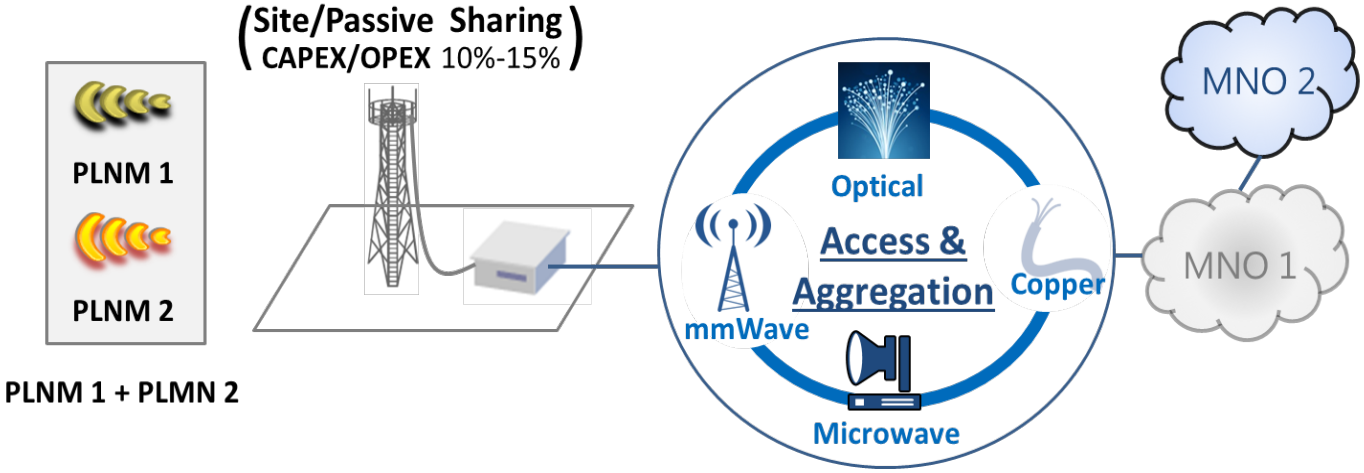




NOF 2017 London

Network Sharing Economics

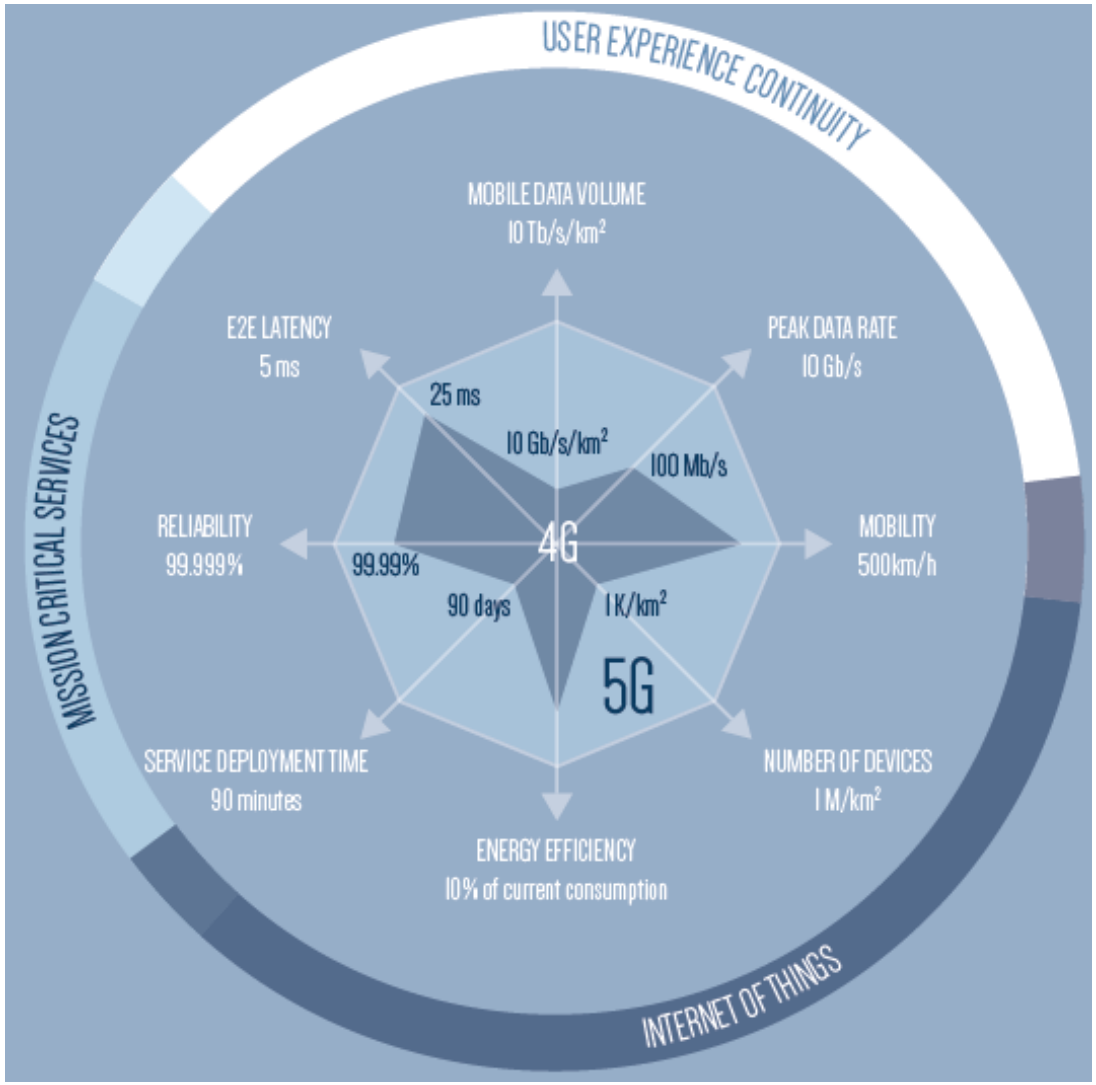
	Radio Access Network (Spectrum and Active Sharing)	Mobile Backhaul (Transport)	Core Network (Evolved Packet Core)
CAPEX	25%-40%	10%-20%	15%-30%
OPEX	20%-30%	10%-15%	20%-25%



- Network sharing reduces CAPEX/OPEX costs
  - for extending coverage, at remote areas
  - for increasing capacity in urban areas
    - no site aquisition - less equipment / cabling
  - providing another revenue source for MNOs

NOF 2017 London

5G in a nutshell

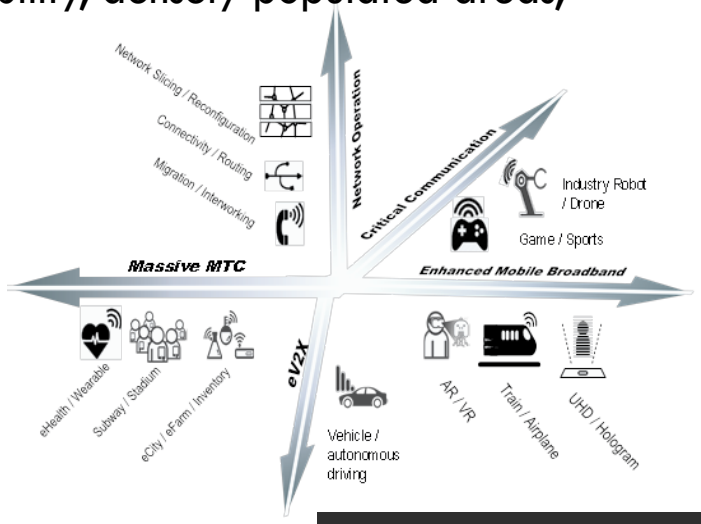


5G Disruptive Capabilities

NOF 2017 London

Study of 5G Networks in 3GPP

- 3GPP completed Stage 1 study – business requirements – on 5G networks (SMARTER)
  - Network operations: Migration, support of flexibility and diverse QoE, converged network access
  - Enhanced mobile broadband: High speed mobility, densely populated areas, ultra-high-definition video (4K), high speed ubiquitous access
  - Critical communications: Delay sensitive (1ms), interactive communications, safety, security
  - Massive MTC: Scalability, energy saving, flexible connectivity
  - Enhanced Vehicular-to-everything: Safety, autonomous driving, onboard multimedia, software upgrades/customization



NOF 2017 London

Multi-tenancy and Network Slicing

Multi-Tenant Emerging Business Players

- Network sharing and multi-tenancy introduces a number of new types of business "players", including the:
  - Infrastructure provider (InP) responsible for the physical network deployment and maintenance. Mobile Network Operators (MNOs) or third parties that interact with other "players" but not with end users directly can take the InP role.
  - Mobile Virtual Network Operator (MVNO) lacks network infrastructure or has limited capacity and/or coverage, and leases resources from an existing InP.
  - Application provider that offers end user applications using a carrier network, - typically OTTs have no control of service quality.
  - Vertical segments/industries that lack network infrastructure but opportunistically or periodically need to reach their customers or enable services orthogonal to the telecommunication industry.

Network Sharing: the way forward

- So far 3GPP has studied
  - **static** network sharing based-on **contractual agreements**
  - concentrating on **Mobile Virtual Network Operators**

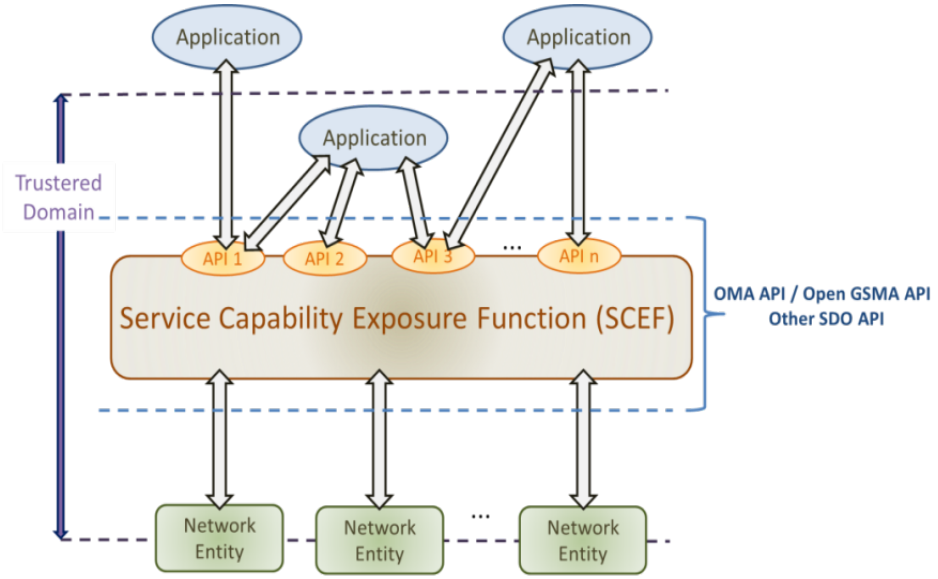
- The revolution towards 5G – Network Slicing
  - Introduce **vertical segments** and **service/application providers**
    - **Assure SLA/QoS: isolation** of different network tenants and service **customization**
  - **On-demand capacity allocation** – including network functions and services
    - flexible sharing on time scales shorter than the contract agreement
  - **Signaling based** resource sharing solutions - **APIs**
    - MNO provide RAN attributes to businesses that lack wireless infrastructure
  - Supporting **a high number of slices in a scalable manner**–
    - Managing user/slice allocation – mobility
  - Economics – **purchase capacity for a specific time** period or purpose

NOF 2017 London

Service Exposure Capacity/Function

- Service Capability Exposure Function (SCEF) located at the operator trust domain
  - securely expose selected service capabilities via network APIs
  - assist 3rd parties to issue network resource request towards MNOs
  - abstracts service capabilities (networking/policy) from underlying network

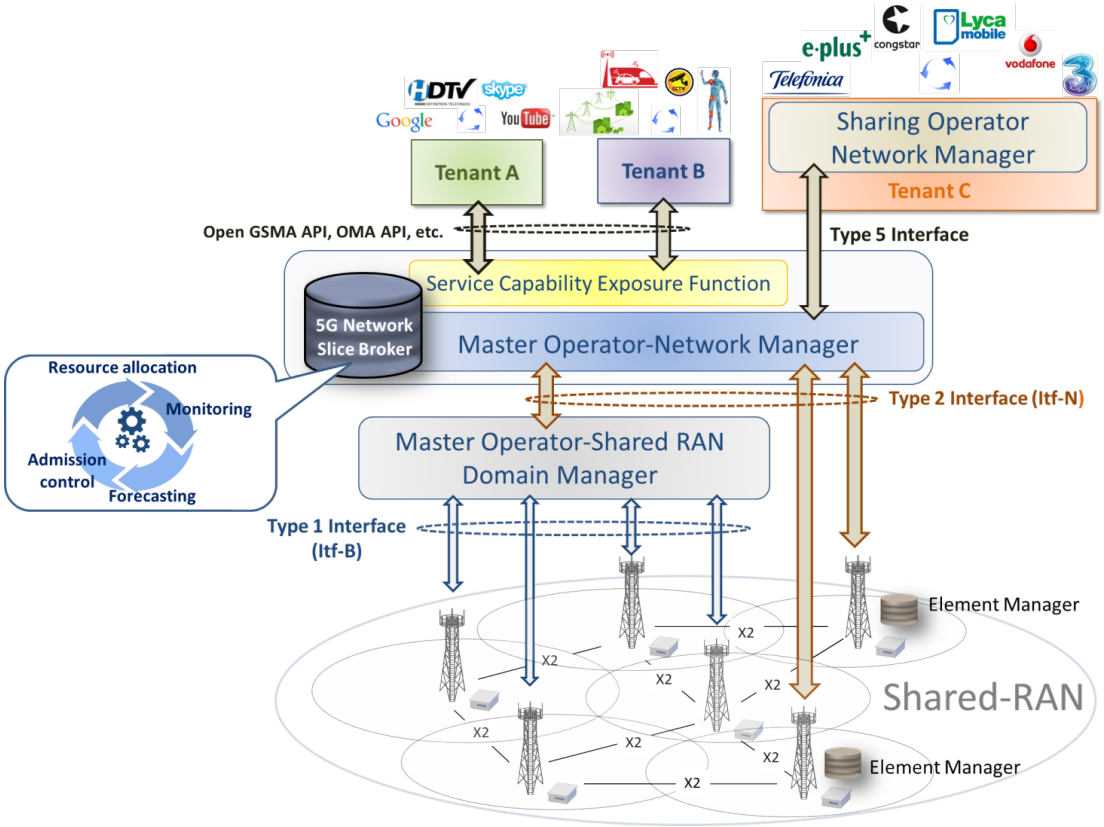
- SCEF is a mediator facilitating:
  - AAA and charging based on offered service and quality provision
  - QoS provision and SLA monitoring for 3rd parties in a dynamic manner
  - user context information
    - location, connectivity, data rate
    - network statue changes
  - provides admission control
    - predictable communication patterns
    - pre-schedule communication timing



NOF 2017 London

Multi-tenancy support in 3GPP Networks

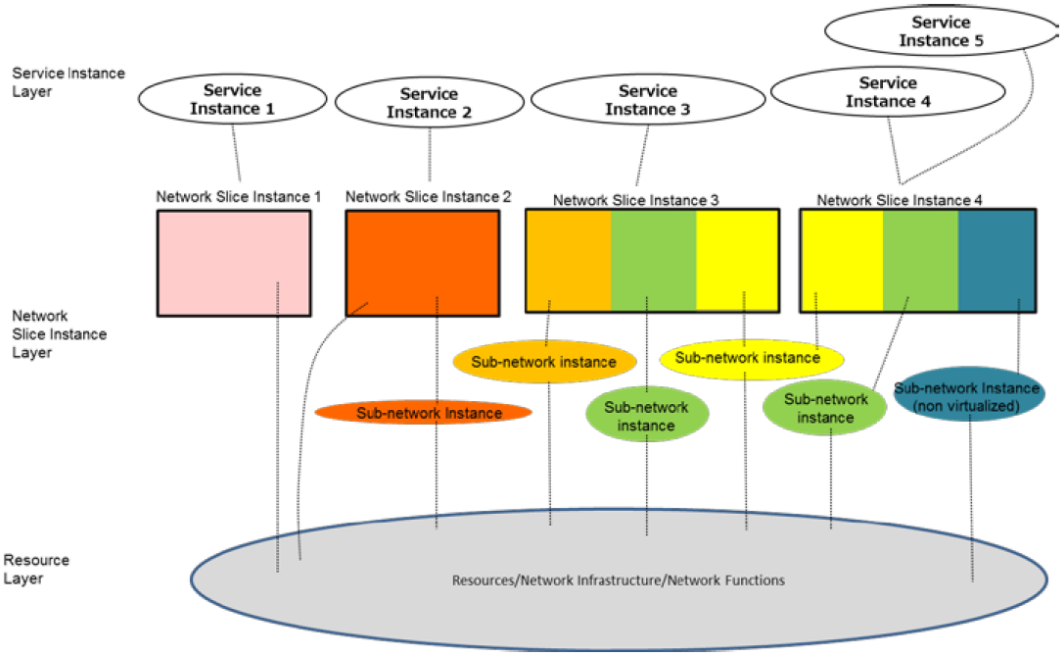
- 5G Network slice broker
  - Perfrom admission control based-on indicated SLA
  - Use Itf-N and Itf-B to monitoring KPIs and configure slice on RAN
  - Receive on-demand slice request from
    - MVNOs via the Type 5.
    - verticals and OTT providers, through SCEF
- Itf-N, Itf-B and Type 5 interface enhancements:
  - PLMN-id or vertical-id
  - type of resources and QoS
  - starting time, duration or periodicity
  - amount of resource blocks or capacity
  - size of file to be downloaded
  - mobility (stationary, low, medium, high)
  - service disruption tolerance
  - performance measurement info



NOF 2017 London

The Network Slicing Concept

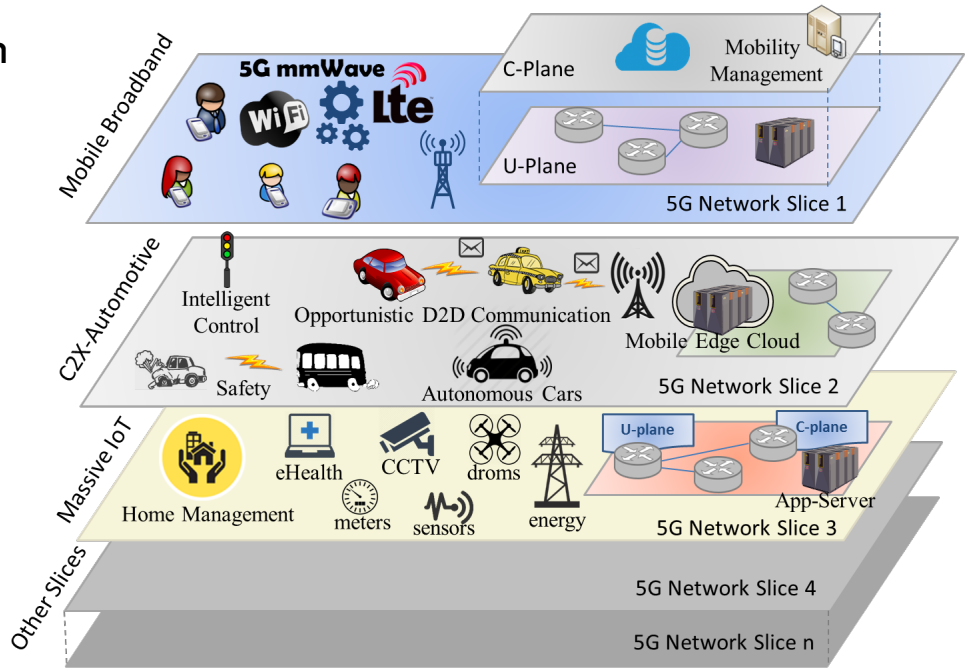
- **Network slicing** – new value creation opportunities(vertical segments)
  - enables a concurrent deployment of **multiple logical**, **self-contained networks** on a common physical infrastructure platform with **devise business demands**
  - offers **dedicated resources** that can be used in an isolated, disjunctive or shared manner and a **customized network operation**
  - supports **flexible, on-demand**, provision of network resources, network functions and applications, **even with short lifecycles** - resource physical or virtual



NOF 2017 London

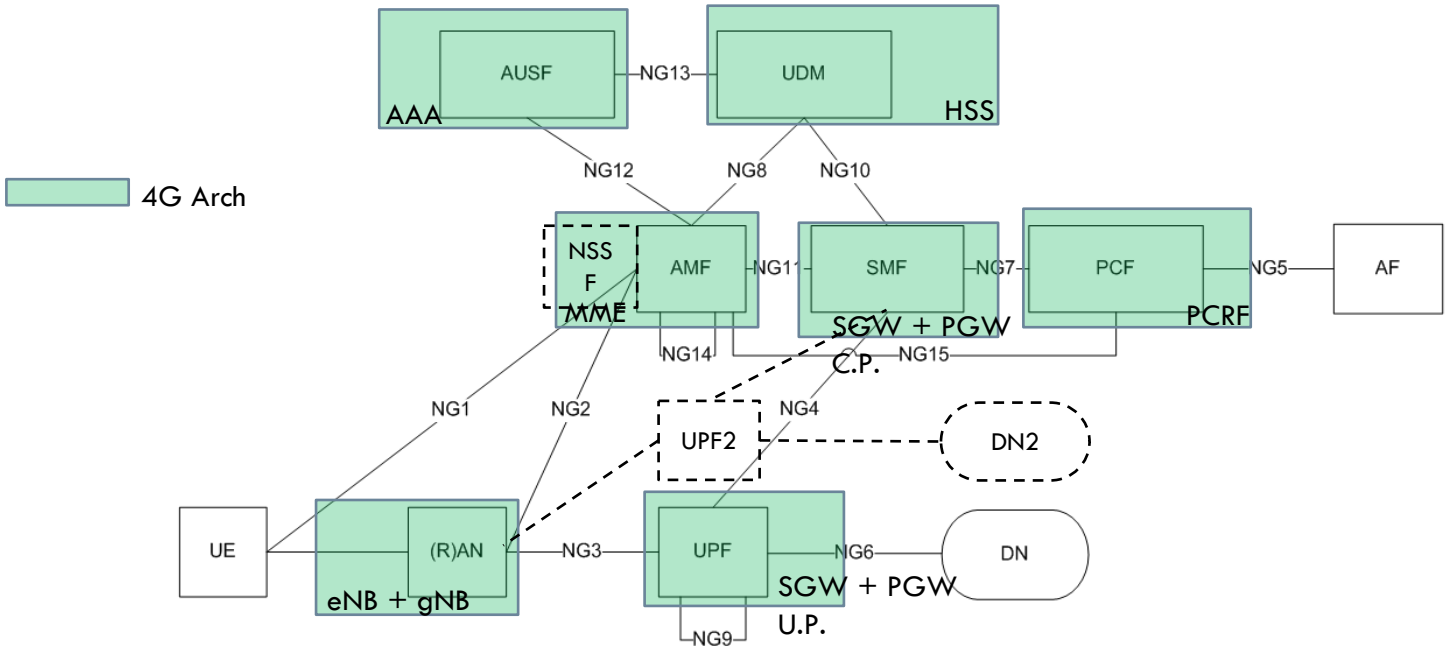
Network Slicing Technology Attributes

- In composing and allocating network slicing:
  - Software defined control and **separation of control/data plane**
    - Network programming via SDN APPs
  - Network function virtualization and resource orchestration
    - **(De)compose/allocate VNFs**
  - **Flexible service chaining** and service provision
  - **Mobile edge computing services** closer to the user
  - QoS provision **policy**
  - **Selection of RAT / fix access**
  - **Data offloading** policies



NOF 2017 London

3GPP Network Slicing Architecture



Network Functions:

-**UDM**: Unified Data Management
-**AUSF**: Authentication Server Function
-**PCF**: Policy Control Function
-**AMF**: Core Access and Mobility Management Function
-**SMF**: Session Management Function
-**UPF**: User plane Function
-**DN**: Data network, e.g., operator services, Internet access
-(*) **NSSF**: Network Slice Selection Function

Interfaces:

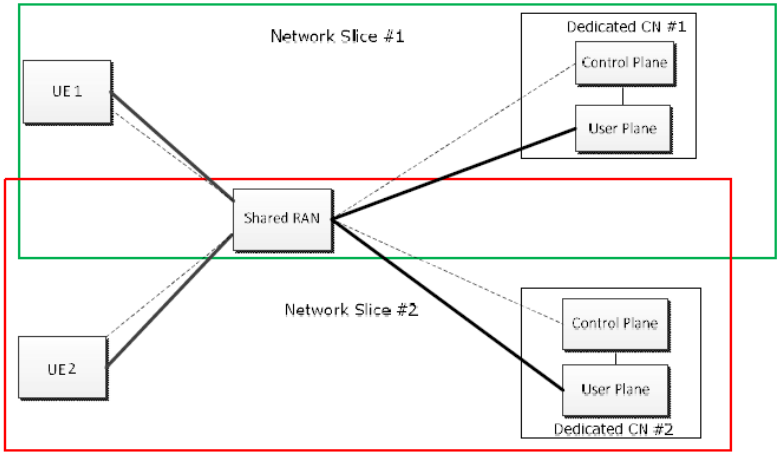
-**NG1**: NAS
-**NG2**: AN-CN C-plane
-**NG3/NG9**: per PDU Session tunnelling
-**NG7/8/10-15**: C-Plane **Service-based interface**

NOF 2017 London

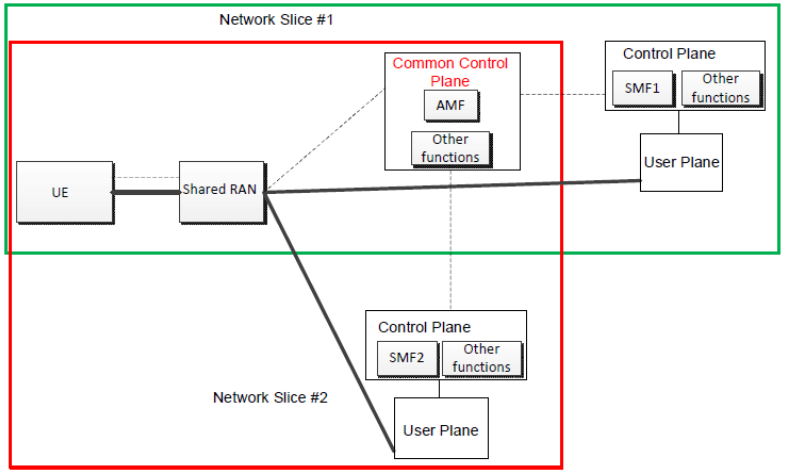
3GPP Network Slicing Architecture

- AMF is the S1 endpoint, where the S1-NAS messages arrive from the UE.
- AMF could be shared between different network slices.
- AMF properly selects the SMF based on the service required.
- UPF is selected accordingly.
- A single tunnel NG3 between between RAN and UPF is established: multiple bearers instantiations will be performed on the same tunnel. UPF marks the messages based on the bearer and the gNB can apply scheduling with different priorities.
- eNB/gNB selects the right AMF based on the service required.
- PCF is also connected to AMF to issue mobility policies.
- SMF and UPF are dedicated per slice (PCF not decided yet).
- **UEs can simultaneously connect to multiple network slices.**

3GPP Network Slicing Example



UE using single slice

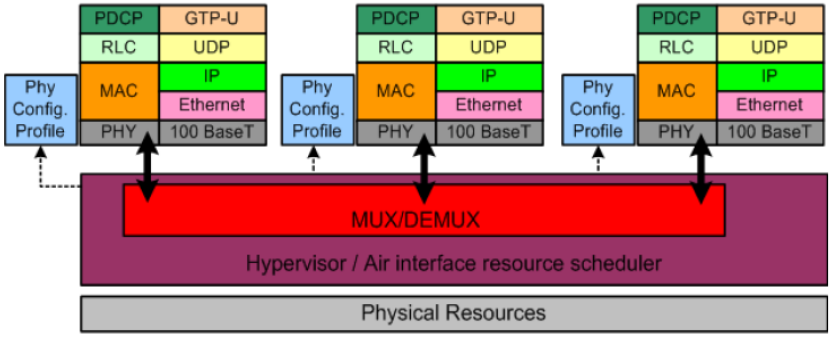


UE using multiple slices

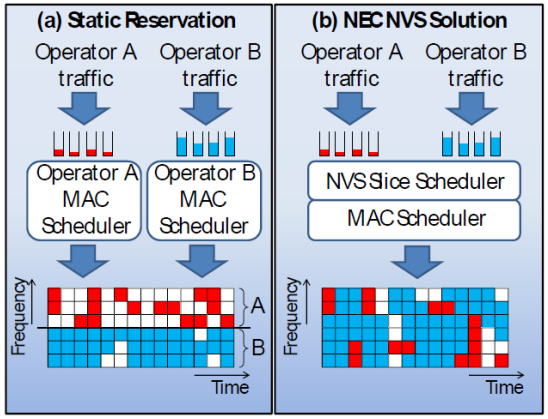
NOF 2017 London

Base Station Virtualization

- eNB virtualization in LTE:
  - via hypervisor means that shares resources among different MNOs
  - consider radio conditions, sharing contracts and traffic load

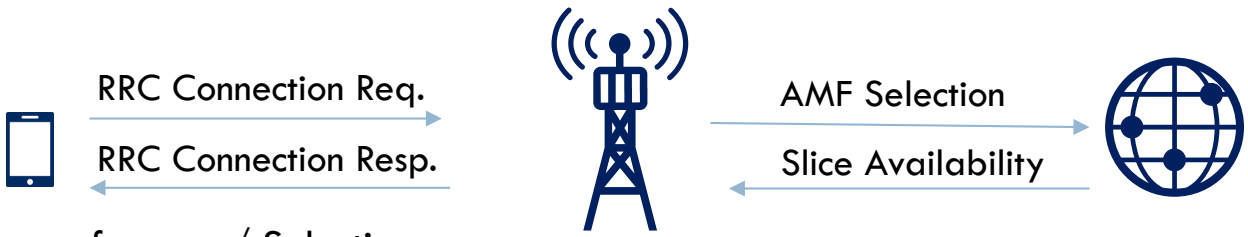


- Network Virtualization Substrate (NVS) operates closely to MAC scheduler and adopts a two-step process:
  - one managed by the infrastructure provider for controlling the resource allocation towards each virtual instance of an eNB
  - the second controlled by each VMNO providing scheduling customization for allocated resources



NOF 2017 London

3GPP RAN for Network Slicing

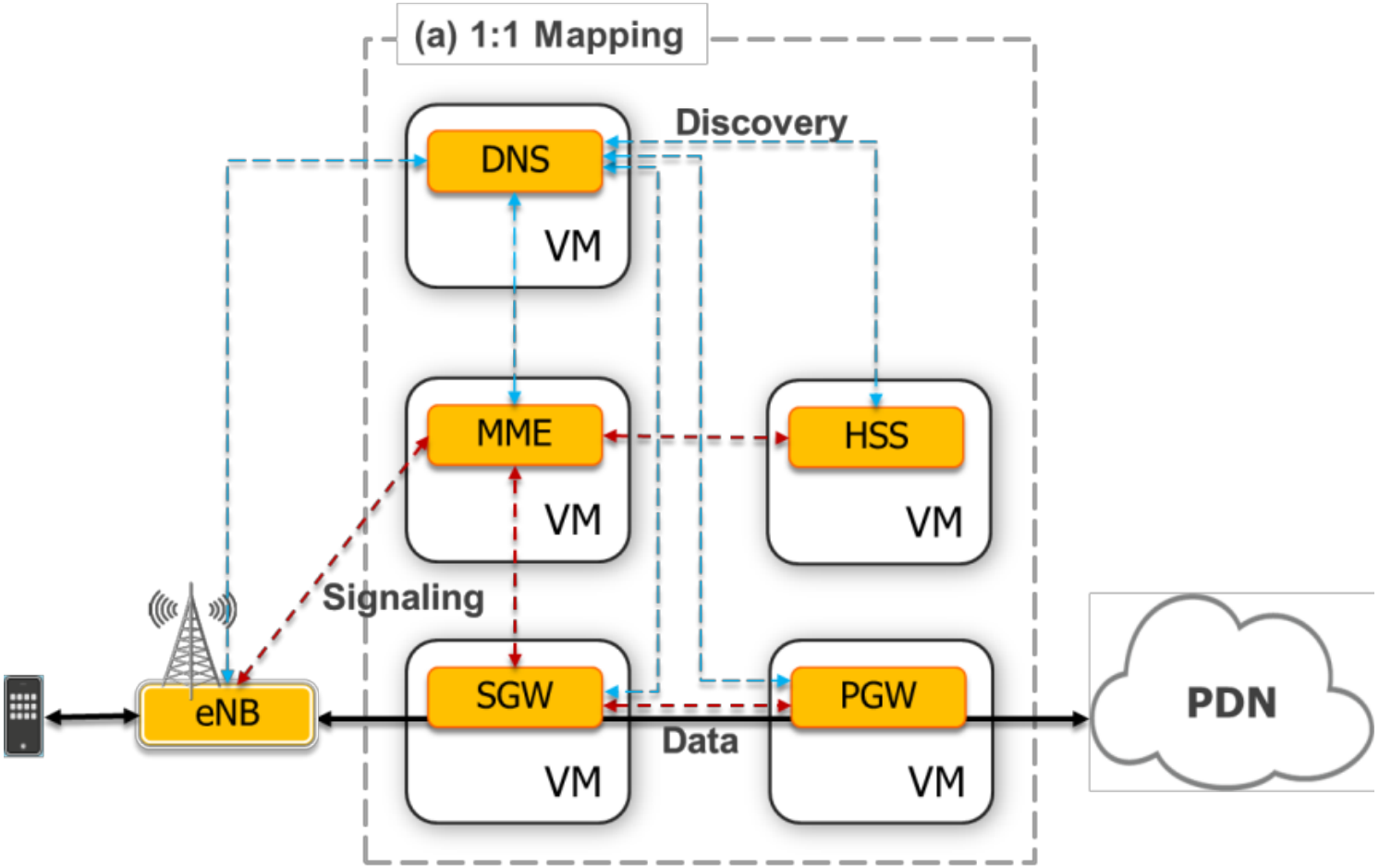


- Slice preference / Selection
  - The UE sends a RRC request specifying the Network Slice Selection Assistance Information (NSSAI) indicator(s).
  - NSSAI has clear definition in SA2 TR 23.799 and TS 23.501.
  - NSSAI is a vector of slice ID preferences (e.g., slice-ids) the UE would like to join.
  - NSSAI information are pre-coded on the SIMcard (like the IMSI information) and cannot be selected (for the moment) by the device user interface.
  - For signaling between RAN and CN a Slice ID is represented by an NSSAI or SM-NSSAI. For the air interface, it is up to RAN groups to decide how to carry NSSAI information in RRC.
  - In the RRC Connection Response, the eNB notifies the UE with allowed/accepted NSSAI, which will be stored and use for future messages.

- Mobility
  - Slice availability during mobility, e.g.,
    - Neighbours may exchange slice availability
    - core network could provide the RAN a mobility restriction list
  - source gNB needs to pass on slices that a UE in question is using to a target gNB
  - target AMF is responsible for removing (or inactivating) at NAS level any slice no longer supported at the target node.

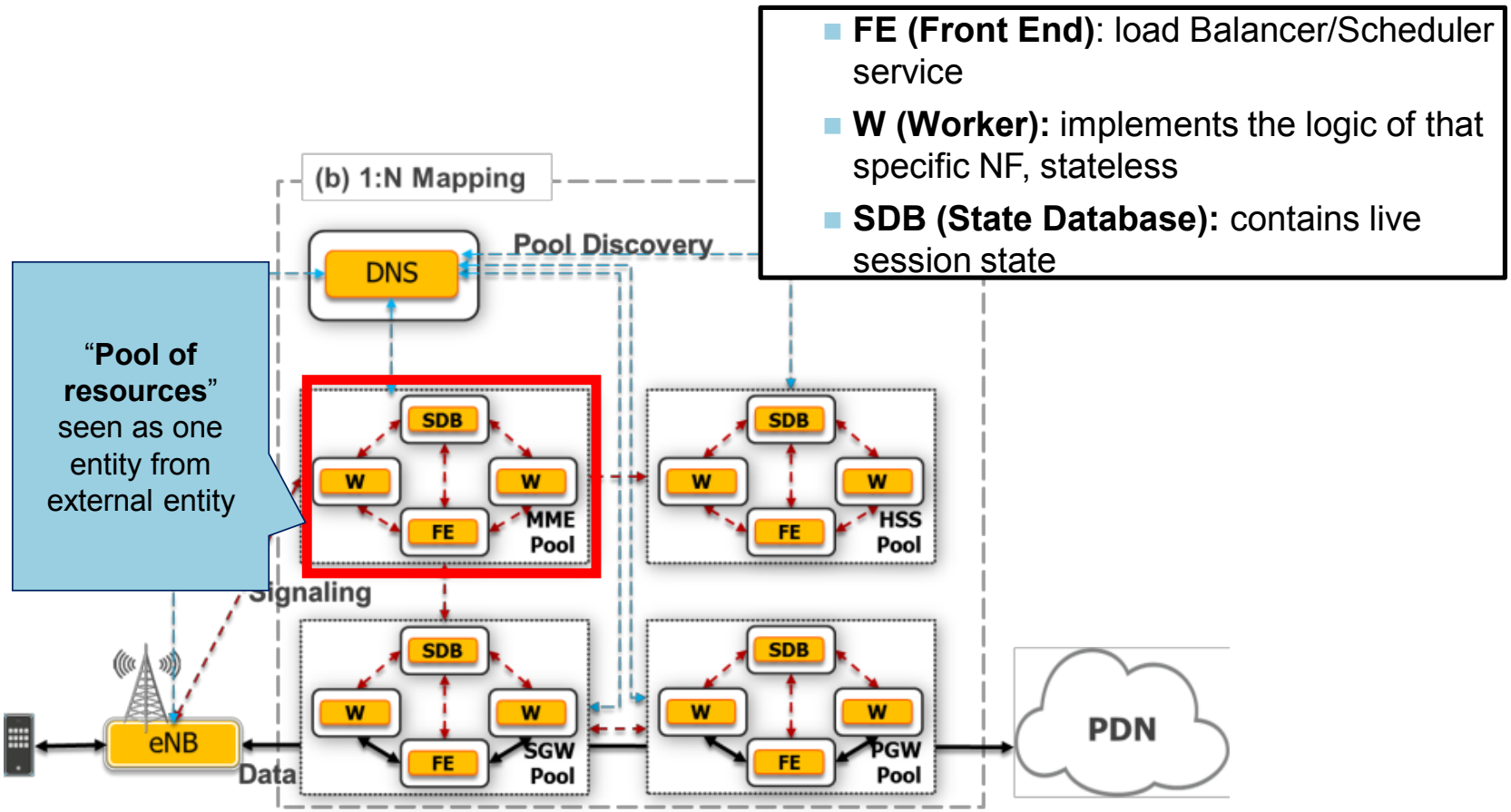
- Heterogeneous Slice Structure
  - Same number of slices available in the same Tracking Area (or specified cell cluster)
  - AMF can deattach the UE from a specific slice, e.g., Netflix when is not used.
  - HO procedures are required when moving to areas with different slice availability, as AMF should be changed.
  - Different PDN-Ips can be assigned to the same UE over time.

EPCaaS



T. Taleb, et al., "EASE: EPC as a Service to Ease Mobile Core Network," IEEE Network Magazine, Vol. 29, No. 2, Mar. 2015.

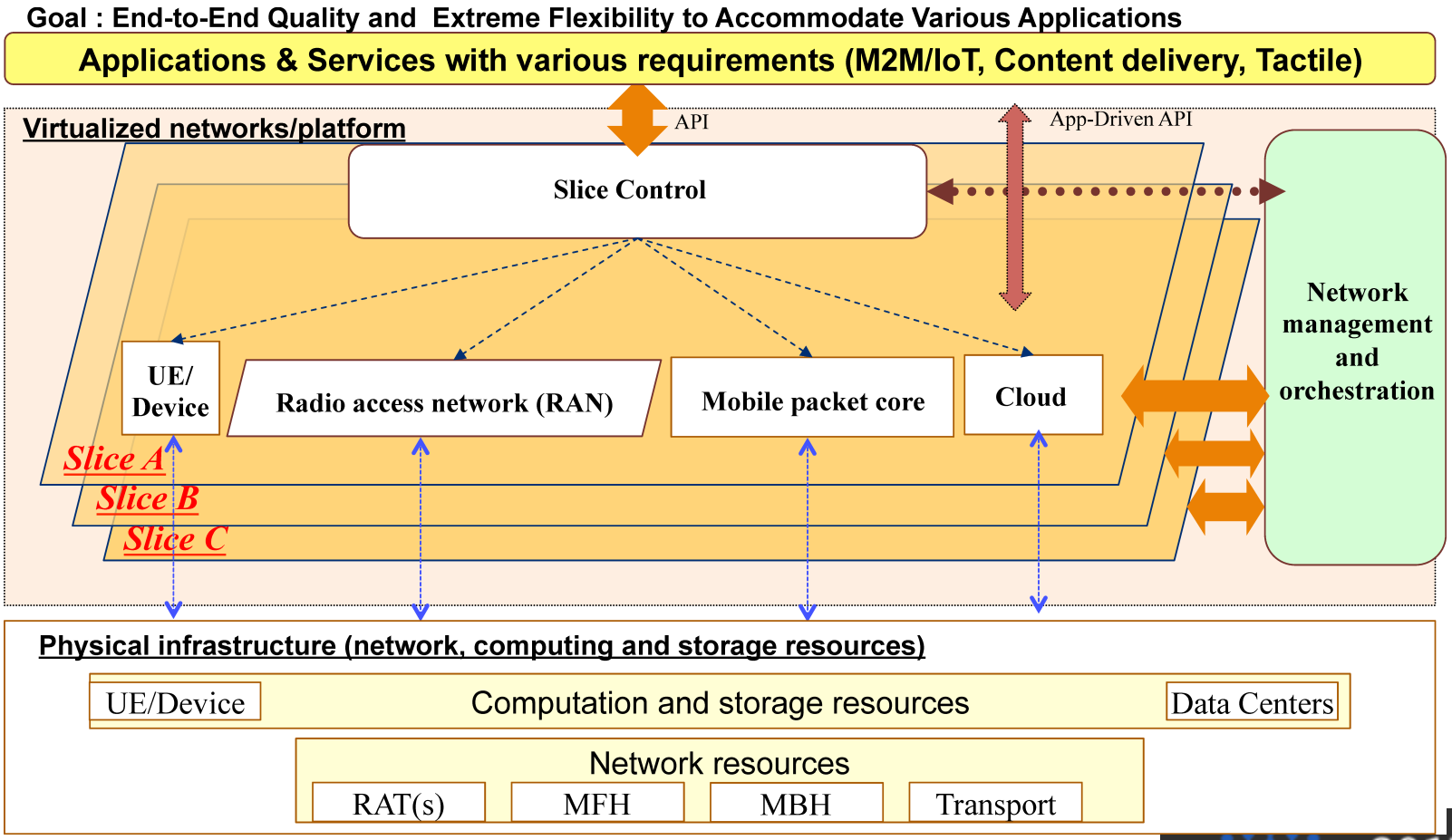
EPCaaS



T. Taleb, et al., "EASE: EPC as a Service to Ease Mobile Core Network," IEEE Network Magazine, Vol. 29, No. 2, Mar. 2015.

NOF 2017 London

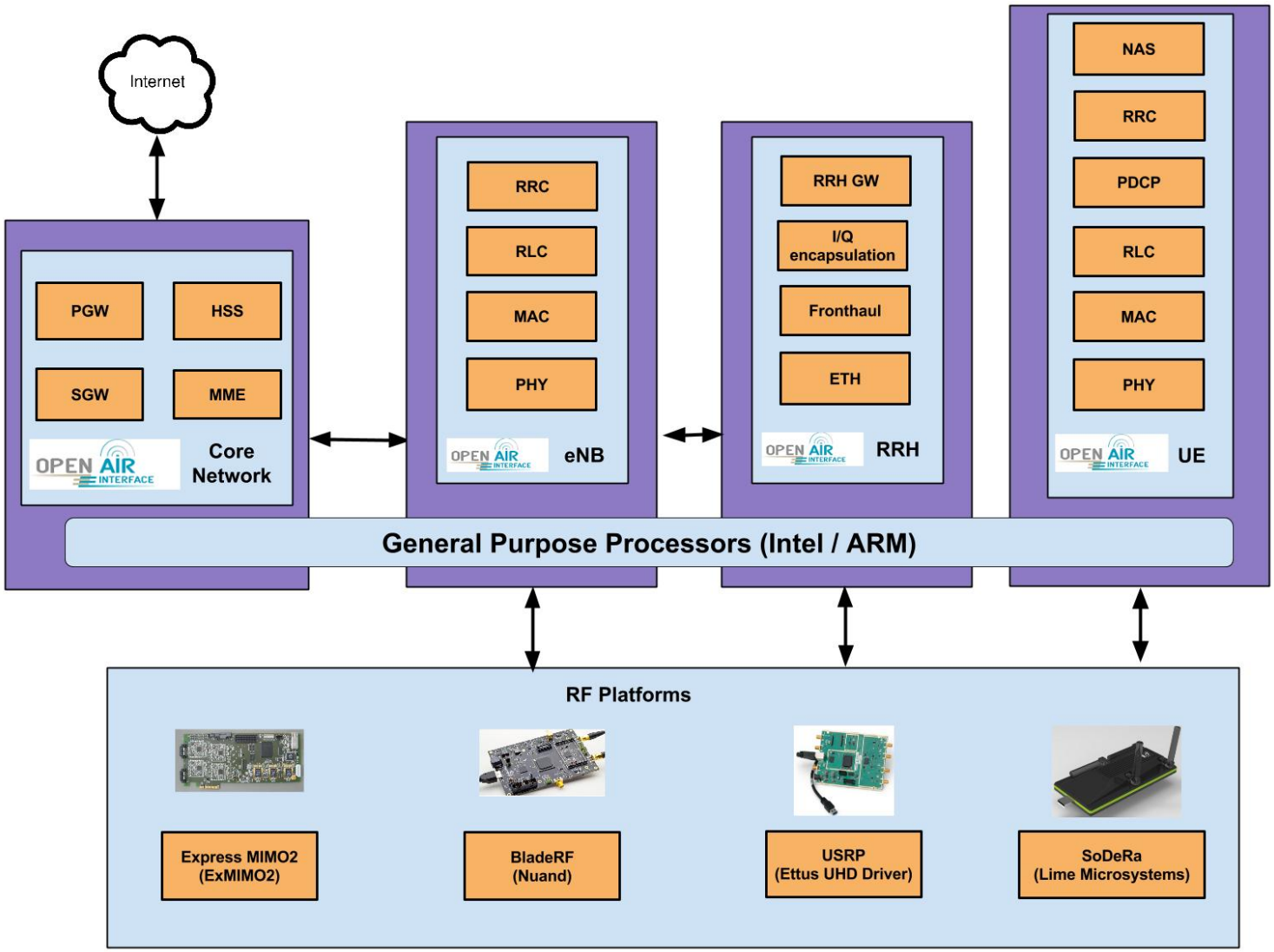
E2E Slicing Concept



A. Nakao, et al., "End-to-end Network Slicing for 5G Mobile Networks," J. Of Information Processing, Vol. 25, pp. 153-163, Feb.'17.

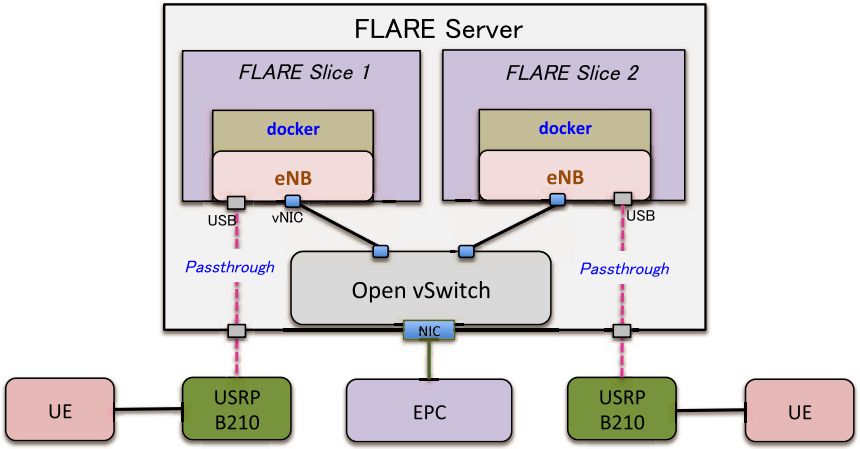
NOF 2017 London

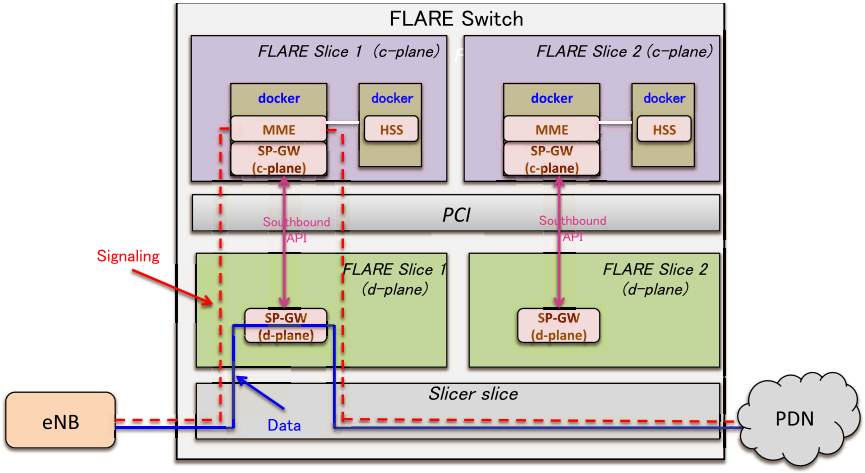
Testbed implementation on Open Air Interface



2017 London

Slicing eNB and EPC

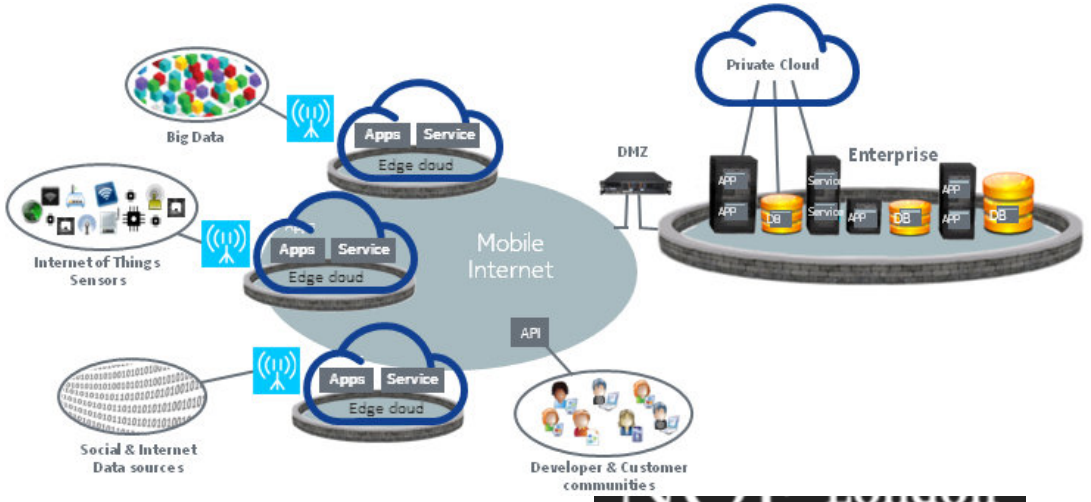




Mobile Edge Computing

- Mobile-edge Computing: A 5G realization
  - offers **IT service environment** and cloud-computing capabilities **within the RAN** in close proximity to mobile subscribers
  - allows content, services and applications to be accelerated, **increasing responsiveness from the edge;** context related services
  - enhanced business exploiting more information about the consumer,
  - **service programmability** greater flexibility for service provisioning

- A 5G realization Service
  - ultra-low latency
  - high-bandwidth
  - exposure to real-time radio network and context info
  - Open edge to 3rd parties



5G Cloud RAN

Cloud RAN - definition

- C-RAN (Cloud-RAN), sometimes referred to as Centralized-RAN, is a proposed architecture for future cellular networks.
- C-RAN is a centralized, cloud computing-based architecture for radio access networks that supports 2G, 3G, 4G and future wireless communication standards.
- Motivation: BTS are designed to handle the maximum traffic, not average traffic, resulting in a waste of processing resources and power at idle times → a more flexible solution is needed.

Evolution of BTS architecture

- All-in-One Macro Base Station
  - In the 1G and 2G cellular networks, base stations had an all-in-one architecture.
  - Analog, digital, and power functions were housed in single cabinet as large as a refrigerator, including supporting facilitates such as power, backup battery, air conditioning, environment surveillance, and backhaul transmission equipment.
  - This all-in-one architecture is mostly found in macro cell deployments.
- Distributed Base Station
  - For 3G, a distributed base station architecture was introduced by Nokia, Huawei and other leading telecom equipment vendors.
  - The radio function unit, also known as the remote radio head (RRH), is separated from the digital function unit, or baseband unit (BBU) by fiber.
  - Digital baseband signals are carried over fiber, using Open Base Station Architecture Initiative (the OBSAI) or Common Public Radio Interface (CPRI) standard.
  - The RRH can be installed on the top of tower close to the antenna, reducing the loss compared to the traditional base station where the RF signal has to travel through a long cable from the base station cabinet to the antenna at the top of the tower.
  - Most modern base stations now use this decoupled architecture.

Evolution of BTS architecture

- C-RAN/Cloud-RAN
  - C-RAN may be viewed as an architectural evolution of the above distributed base station system.
  - It takes advantage of many technological advances in wireless, optical and IT communications systems.
    - It uses the latest CPRI standard, low cost Coarse or Dense Wavelength Division Multiplexing (CWDM/ DWDM) technology, and mmWave to allow transmission of baseband signal over long distance, thus achieving large scale centralised base station deployment.
  - It applies recent Data Centre Network technology to allow a low cost, high reliability, low latency and high bandwidth interconnect network in the BBU pool.
  - It utilises open platforms and real-time virtualisation technology rooted in cloud computing to achieve dynamic shared resource allocation and support of multi-vendor, multi-technology environments.

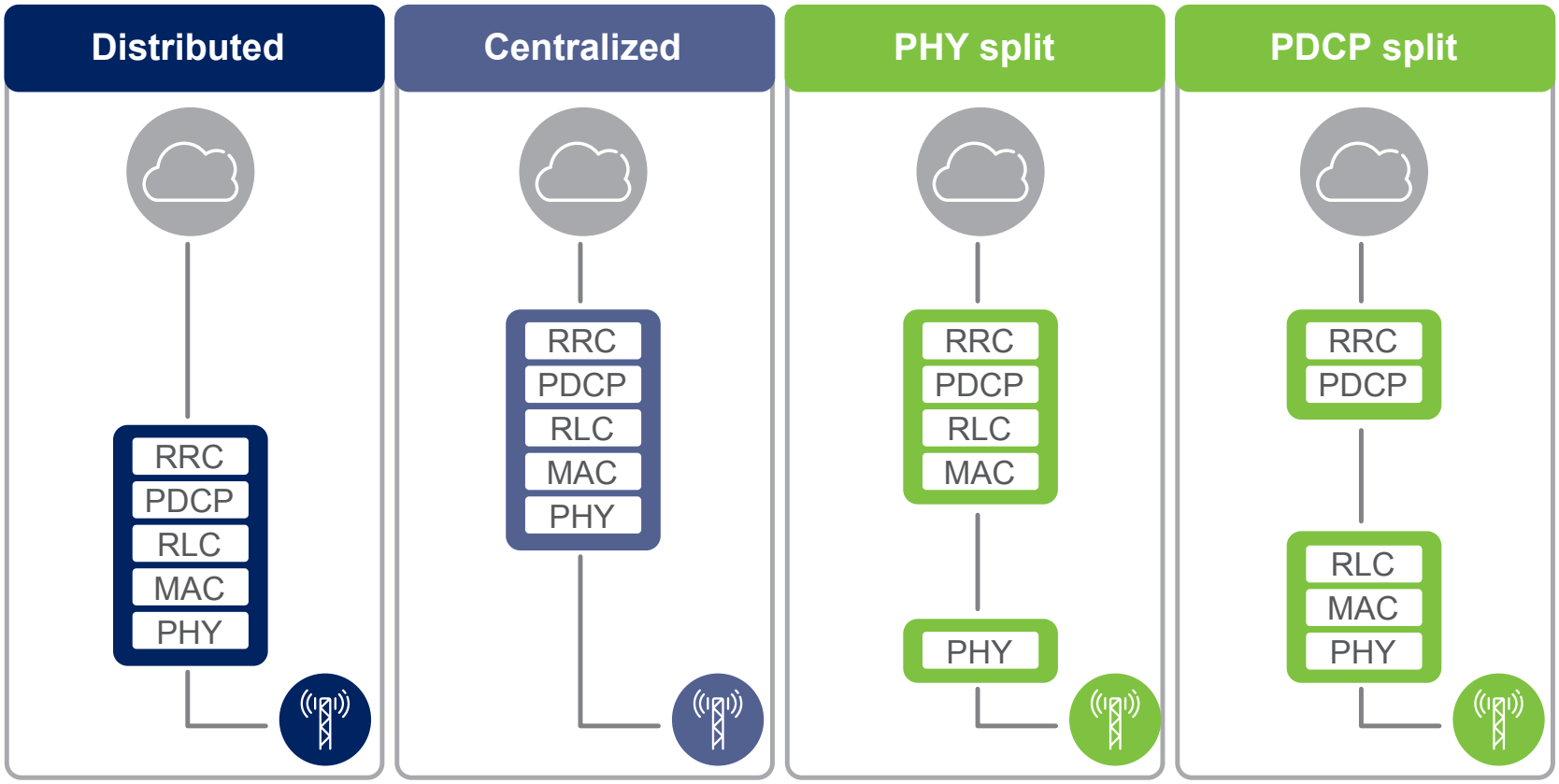
Cloud RAN Architecture

- Large scale centralized deployment:
  - It allows hundreds of thousands of remote RRH connect to a centralized BBU pool. The maximum distance can be 20 km in fiber link for 4G (LTE/LTE-A) system, even longer distance (40 km~80 km) for 3G (WCDMA/TD-SCDMA) and 2G (GSM/CDMA) systems.
  - Apparently, some Asia operators have deployments of C-RAN systems with 1200 RRHs centralized to one central office.
- Native support to Collaborative Radio technologies:
  - Any BBU can talk with other BBU within the BBU pool with very high bandwidth (10Gbit/s and above) and low latency (10us level). This is enabled by the interconnect of BBU in the pool.
  - This is one major difference from BBU Hoteling, or base station hoteling. In the later case, the BBU of different base stations are simply stacked together and has no direct link among them to allow physical layer co-ordination.

Cloud RAN Architecture

- Real-time virtualization capability based on open platform:
  - This is different from the traditional base station built on proprietary hardware, where the software and hardware are closed-sources and provided by one single vendor. C-RAN BBU pool is built on open hardware, like x86/ARM CPU based servers, plus interface cards to handle fiber link to RRH and inter-connection in the pool.
  - Real-time virtualization make sure the resources in the pool can be allocated dynamically to base station software stacks, say 4G/3G/2G function modules from different vendors according to network load.
  - To satisfy the strict timing requirement of wireless communication system, the real-time performance for C-RAN is at the level of 10s of micro-seconds, which is two magnitude higher than the milli-second level 'real-time' performance usually seen in Cloud Computing environment.

Cloud RAN Functional Split



See Ericsson, "Cloud RAN" White Paper, Uen 284 23-3271, Sept. 2015.

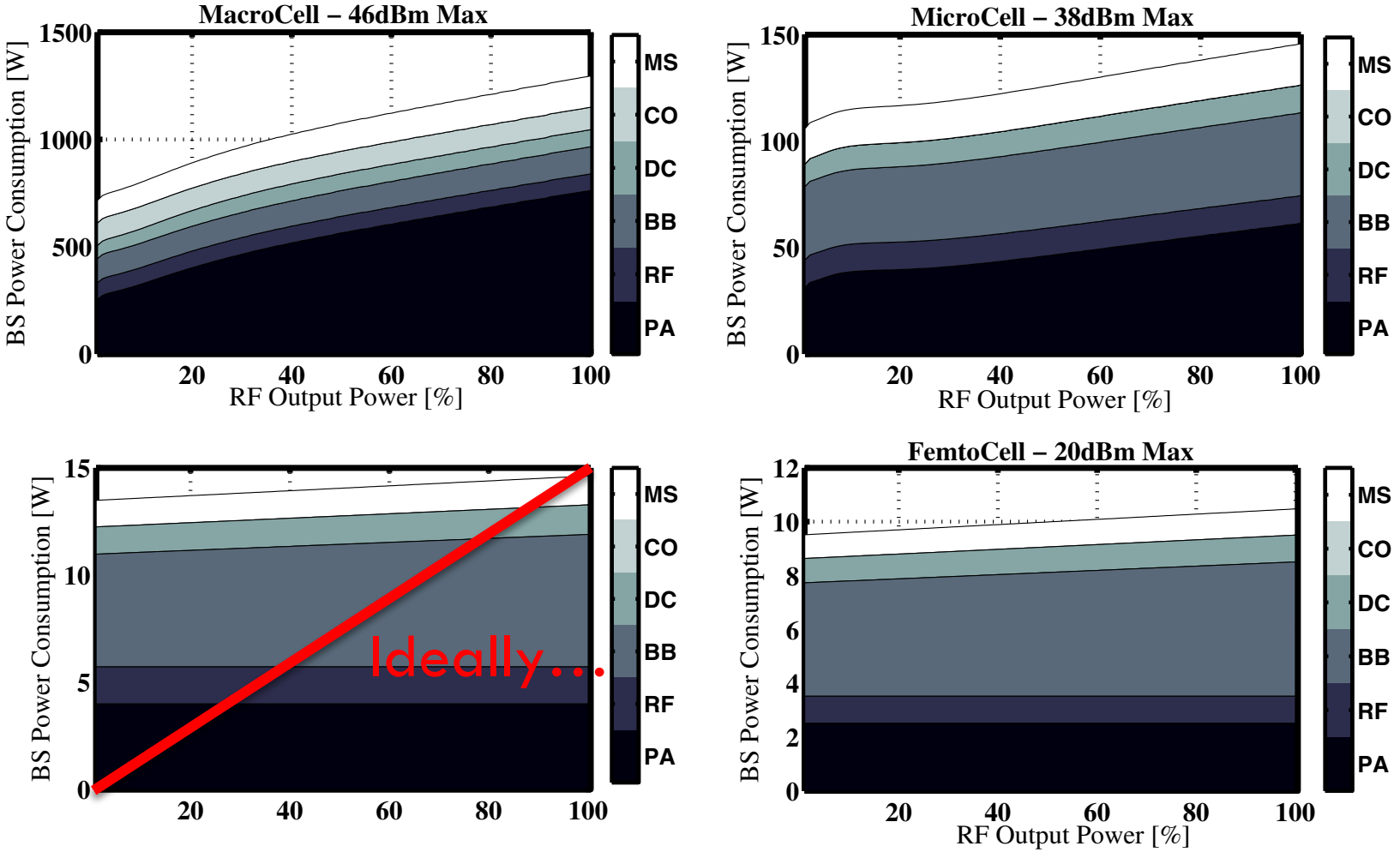
Typical Power Consumption (heterogeneous networking)

LTE Base Station Power Consumption at Maximum Load for Different BS Types as of 2010.

			Macro	Micro	Pico	Femto
PA	$P_{\max}$	[dBm]	46.0	38.0	21.0	17.0
		[W]	40.0	6.3	0.13	0.05
	Back-off	[dB]	8.0	8.0	12.0	12.0
	PA Efficiency	[%]	31.1	22.8	6.7	4.4
	Total PA, P_{PA}	[W]	**128.2**	**27.7**	**1.9**	**1.1**
RF	P_{TX}	[W]	6.8	3.4	0.4	0.2
	P_{RX}	[W]	6.1	3.1	0.4	0.3
	Total RF, P_{RF}	[W]	**13.0**	**6.5**	**1.0**	**0.6**
BB	Radio (inner Rx/Tx)	[W]	10.8	9.1	1.2	1.0
	Turbo code (outer Rx/Tx)	[W]	8.8	8.1	1.4	1.2
	Processors	[W]	10.0	10.0	0.4	0.3
	Total BB, P_{BB}	[W]	**29.5**	**27.3**	**3.0**	**2.5**
DC-DC, σ_{DC}		[%]	7.5	7.5	9.0	9.0
Cooling, σ_{cool}		[%]	10.0	0.0	0.0	0.0
Mains Supply, σ_{MS}		[%]	9.0	9.0	11.0	11.0
Total per TRX chain		[W]	**225.0**	**72.3**	**7.3**	**5.2**
# Sectors		#	3	1	1	1
# Antennas		#	2	2	2	2
# Carriers		#	1	1	1	1
Total N_{TRX} **chains,**	P_{in}	[W]	**1350.0**	**144.6**	**14.7**	**10.4**

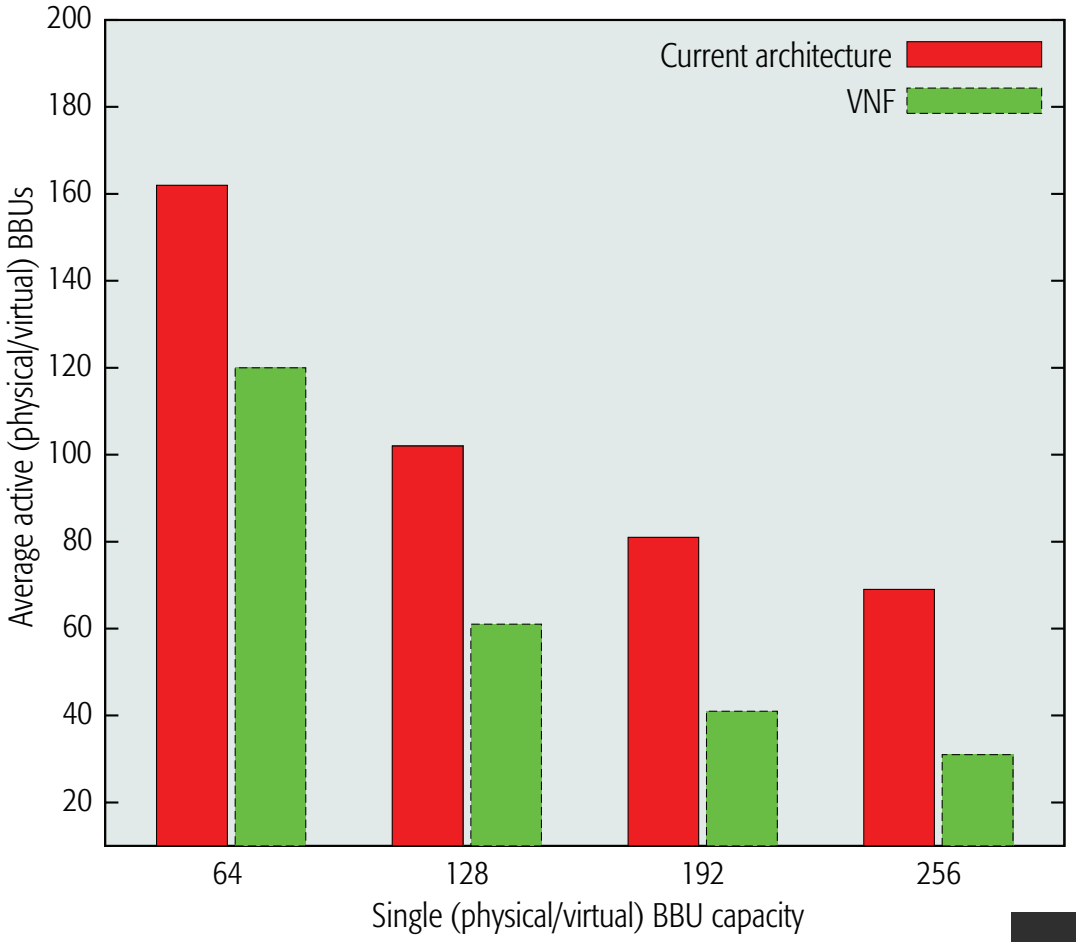
Gunther Auer, et al., "Cellular Energy Efficiency Evaluation Framework," IEEE VTC 2011.

Typical Power Consumption (heterogenous networking)



Gunther Auer, et al., "Cellular Energy Efficiency Evaluation Framework," IEEE VTC 2011.

Cloud RAN: quantitative benefits?

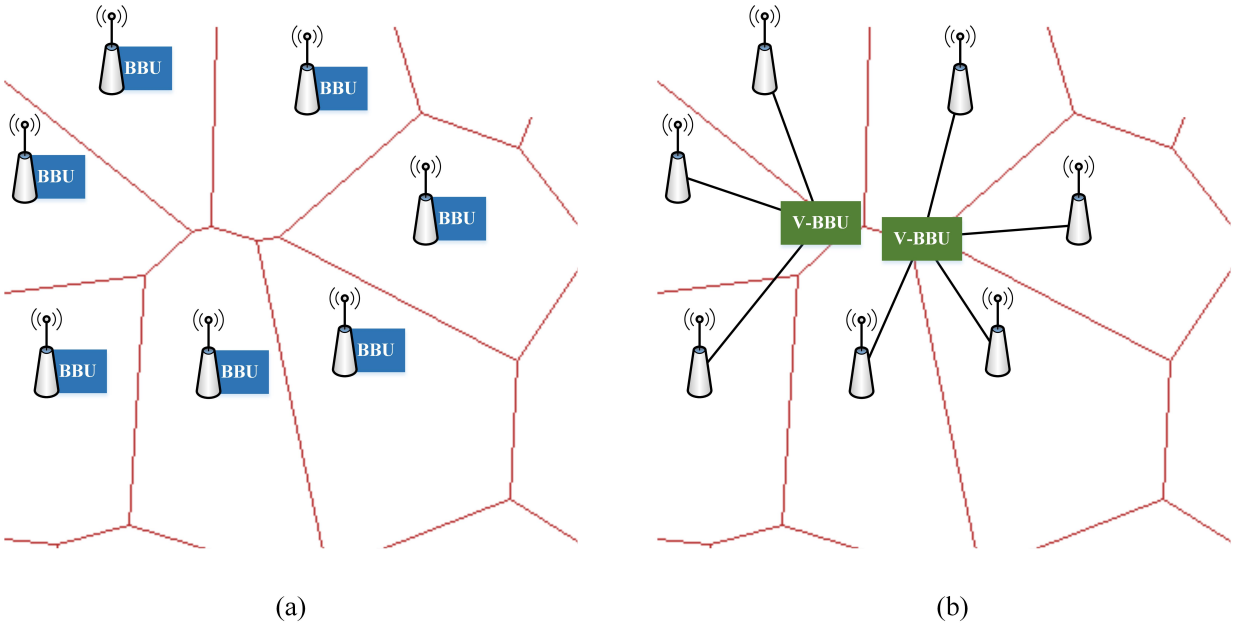


S. Abdelwahab, B. Hamdaoui, M. Guizani, T. Znati, "Network Function Virtualization in 5G", IEEE COMMAG, 2016.

NOF 2017 London

Cloud RAN modeling using stochastic geometry [*]

- To define a quantative model for analyzing the potential benefits of Cloud RAN

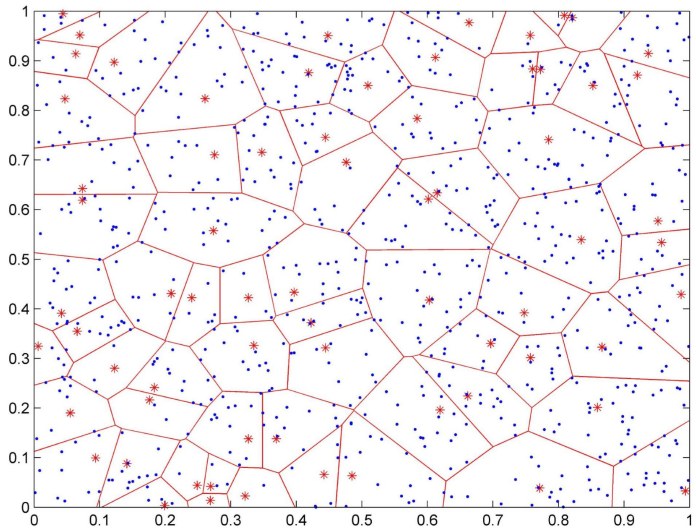


[*] R. Bassoli, F. Granelli, M. Di Renzo, "Energy Efficient Design of 5G Cloud Radio Access Network," in preparation.

NOF 2017 London

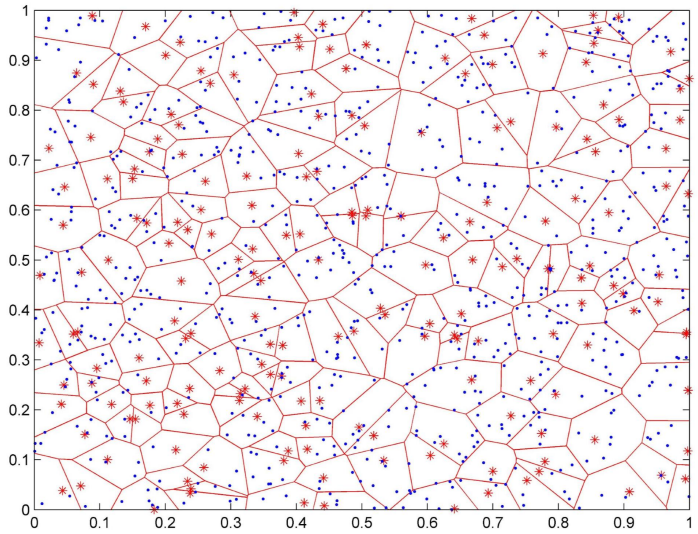
Cloud RAN modeling using stochastic geometry

- Stochastic geometry is employed to study the coverage requirements and optimal AP locations



(a)

Micro-cells

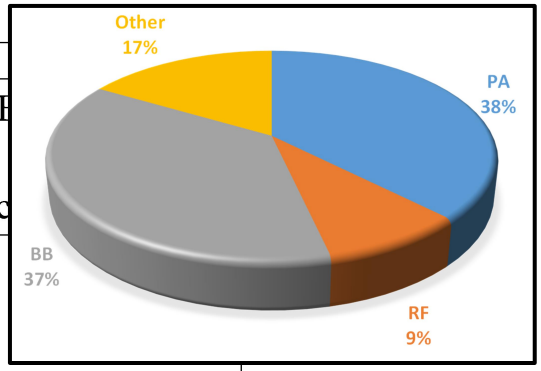


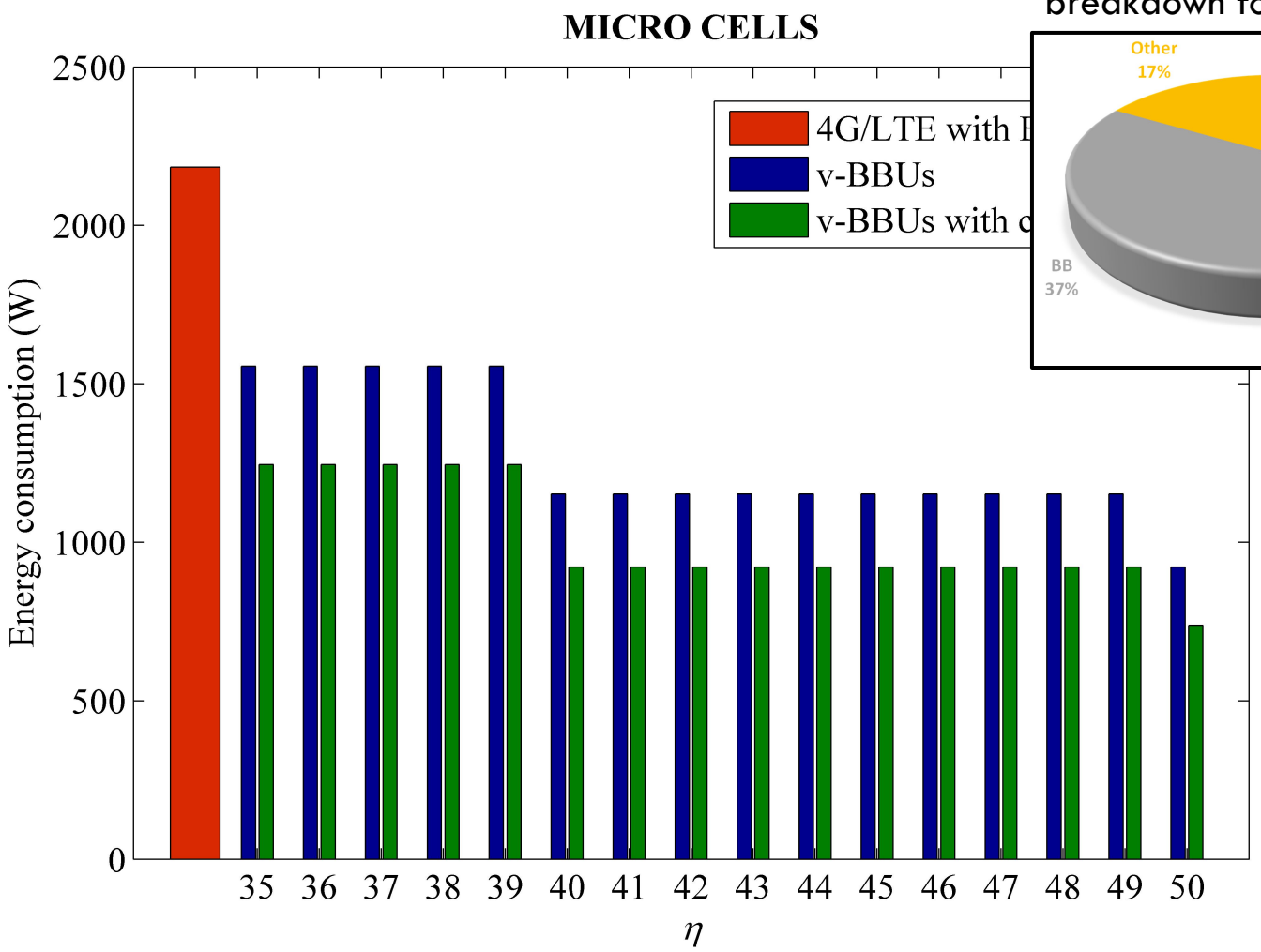
(b)

Pico-cells

Premilinary results

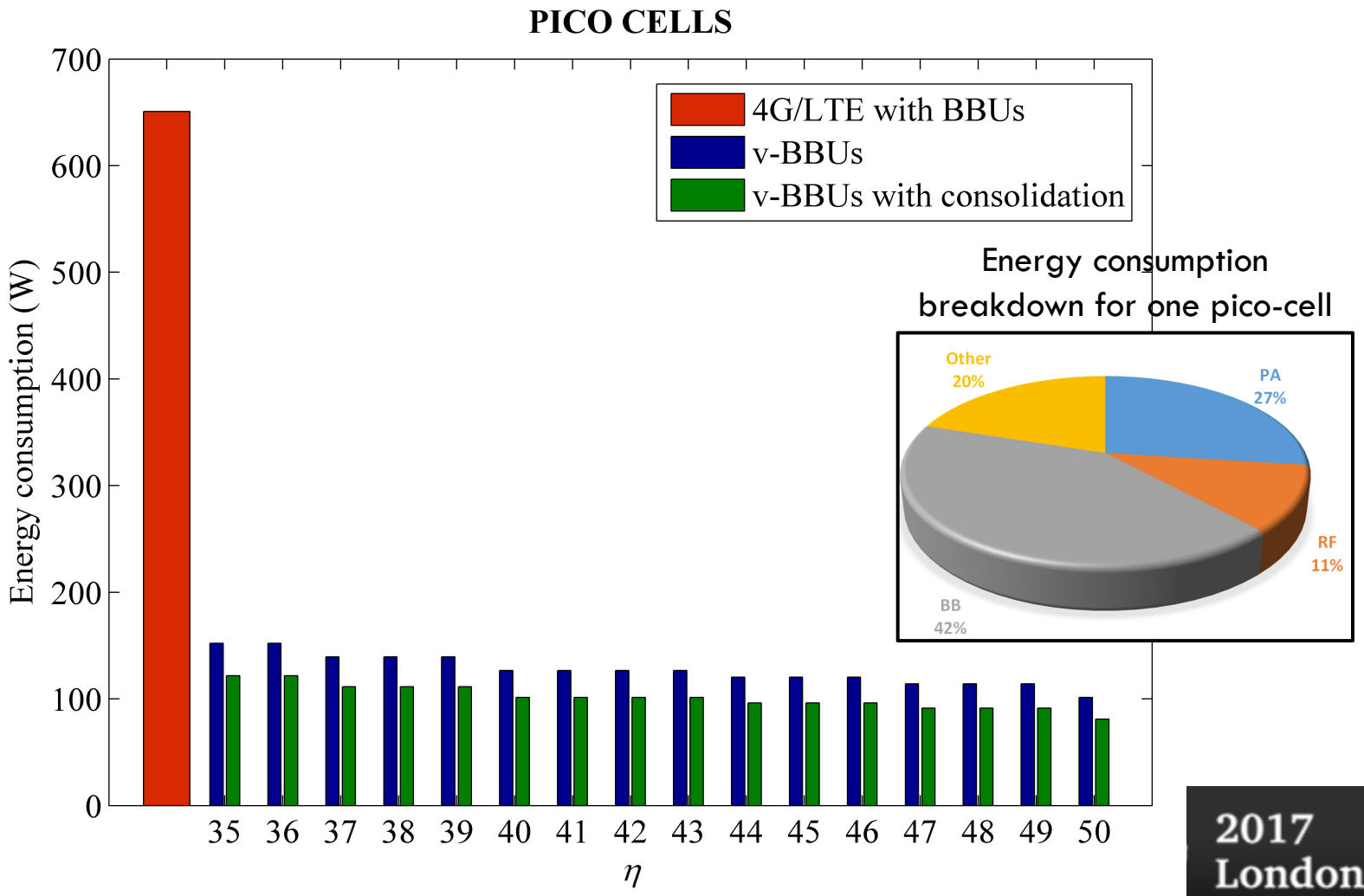
Energy consumption breakdown for one micro-cell





2017 London

Premilinary results



2017
London

Conclusions

- The tutorial provided an introduction to the network softwarization and virtualization technologies
- Such technologies are then applied to 5G architecture design for enabling network slicing and Cloud RAN
- Ongoing standardization is considering those technologies, possibly for later releases of 5G 3GPP standard

Any questions?

Fabrizio Granelli

fabrizio.granelli@unitn.it