UNC CHARLOTTE

Learning-Assisted Network Slicing for Diverse Applications in 5G

Tao Han

The Department of Electrical and Computer Engineering

The University of North Carolina at Charlotte, NC, United States

Tao.Han@uncc.edu

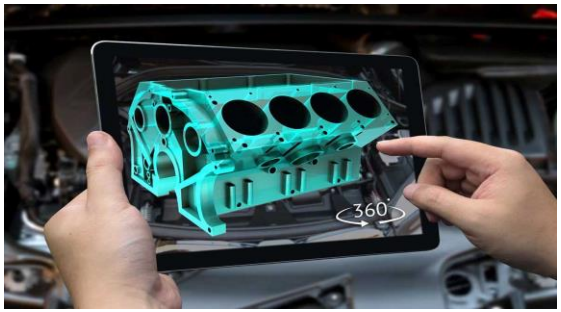
https://webpages.uncc.edu/than3/index.html

What are “killer” applications for 5G?

Autonomous Vehicle

Smart Industry

Gaming



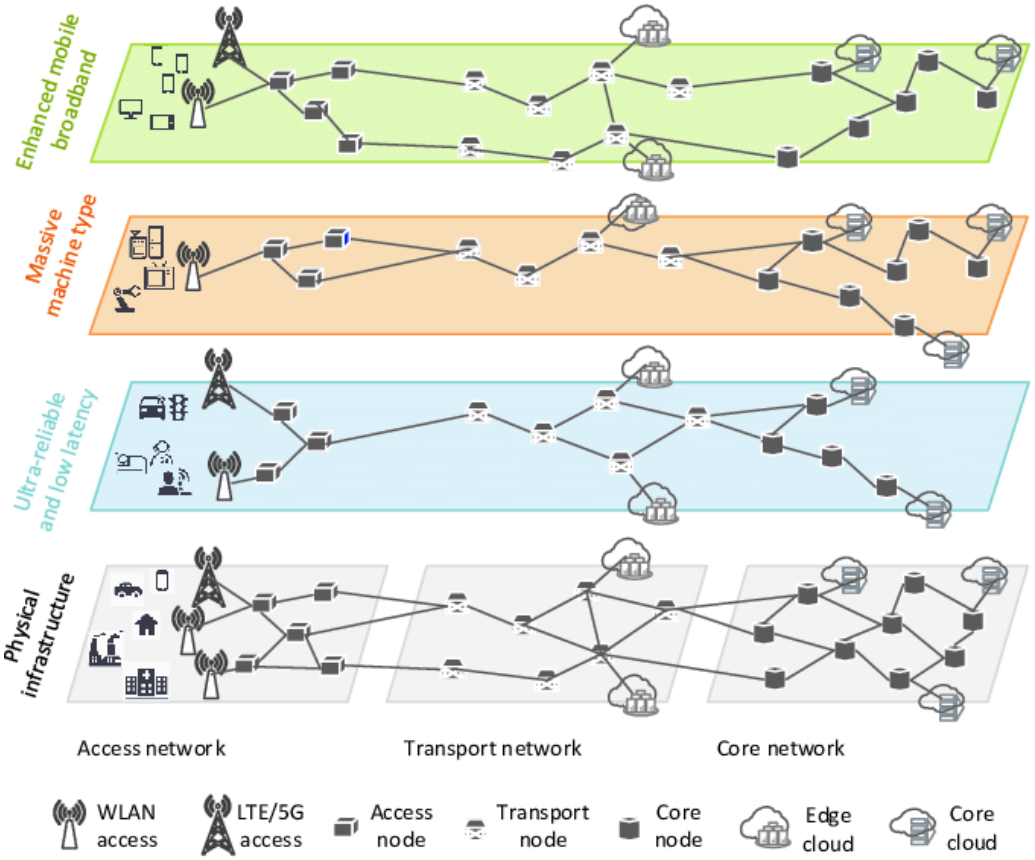
Education

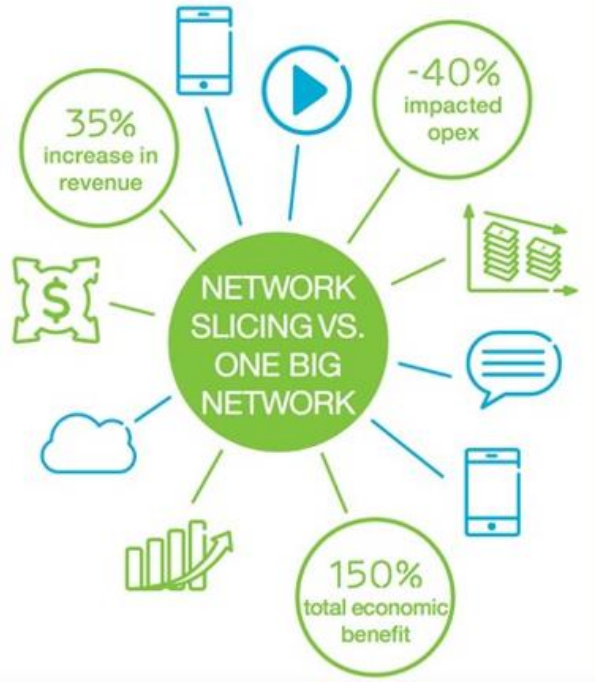
Battlefield

Tourist & Navigation

- Diverse resource requirements from multiple domains:
  - Radio Access Networks
  - Transportation
  - Computing

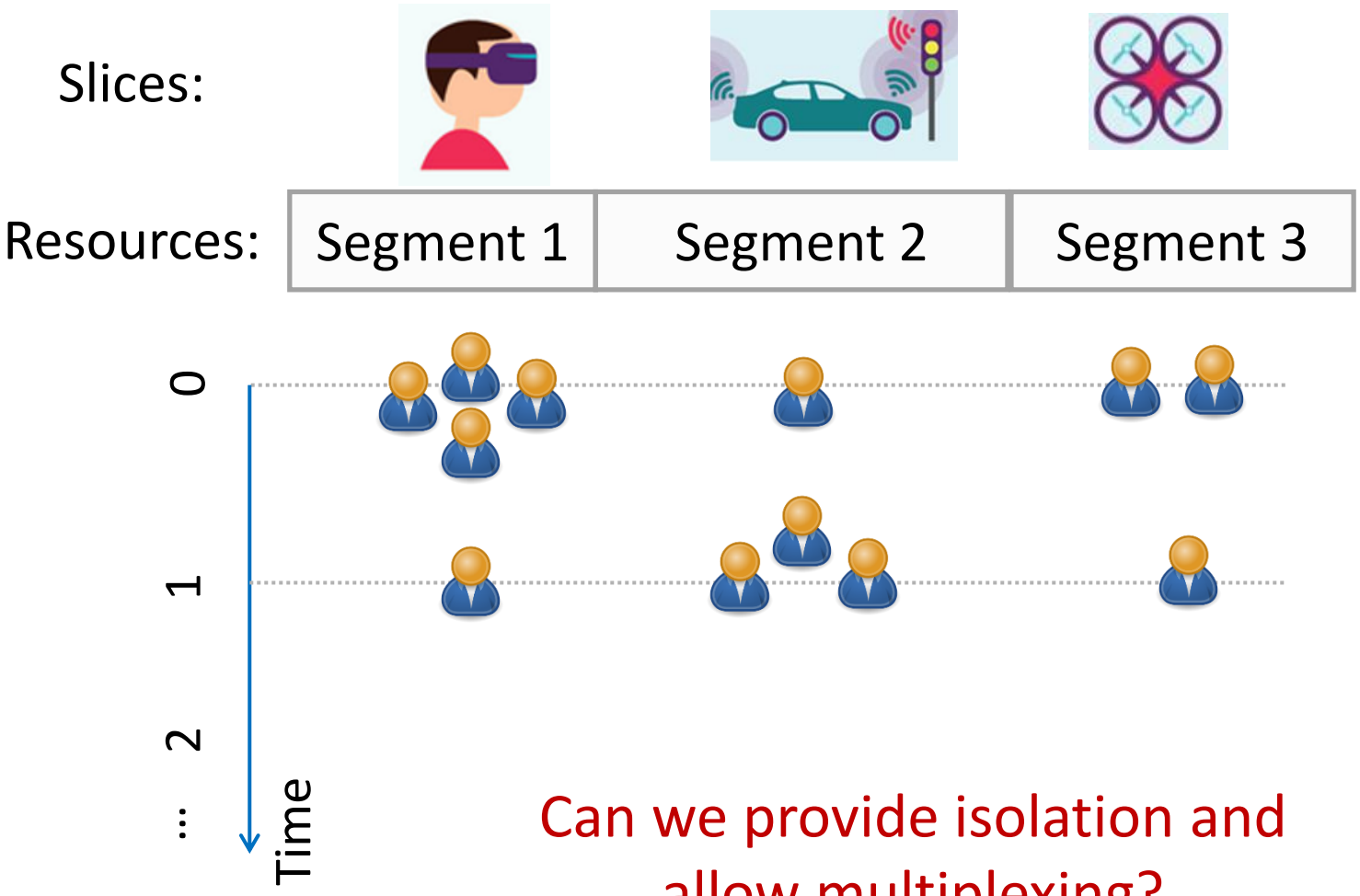
Network Slicing: End-to-End Customization





* Guan, W., Wen, X., Wang, L., Lu, Z. and Shen, Y., 2018. A service-oriented deployment policy of end-to-end network slicing based on complex network theory. IEEE Access, 6, pp.19691-19701.

Isolation v.s. Multiplexing



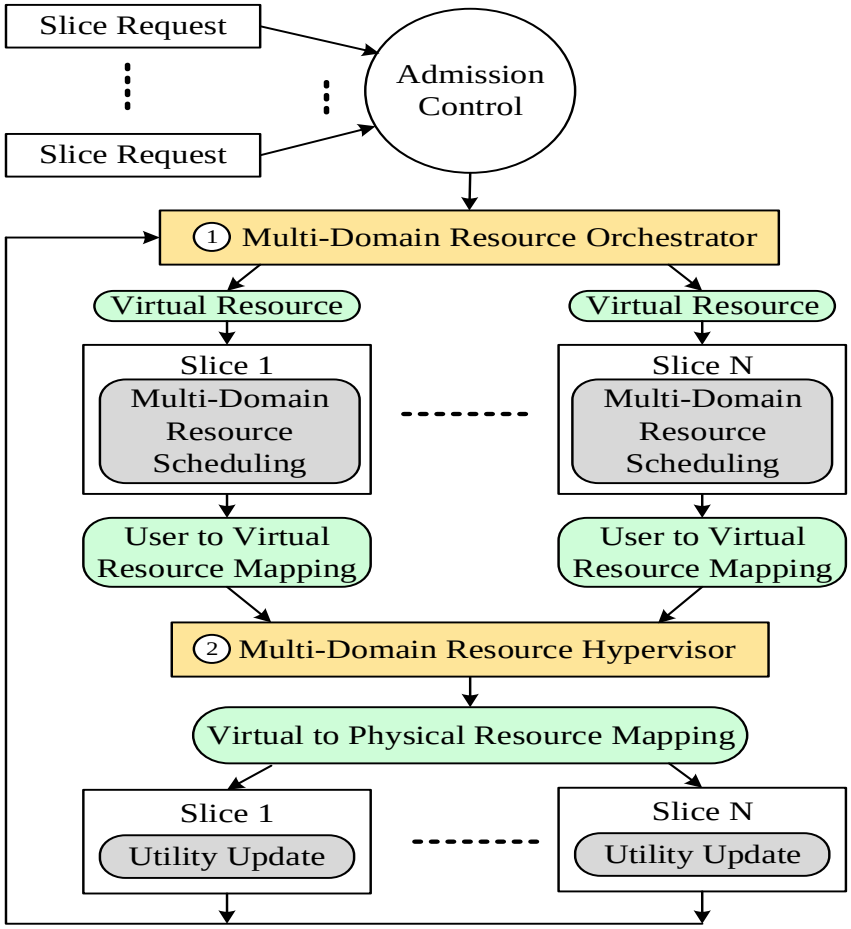
What is the “Capacity” formula?

- The Shannon–Hartley theorem

$$C = B\ log_2(1 + SNR)$$

	RAN	Transport	Edge/Cloud Computing
20 ms	Good Better Best ✓	Good Better Best ✓	Good Better Best ✓
200 ms	Good Better ? Best	Good Better ? Best	Good Better ? Best

Learning-Assisted Dynamic Network Slicing

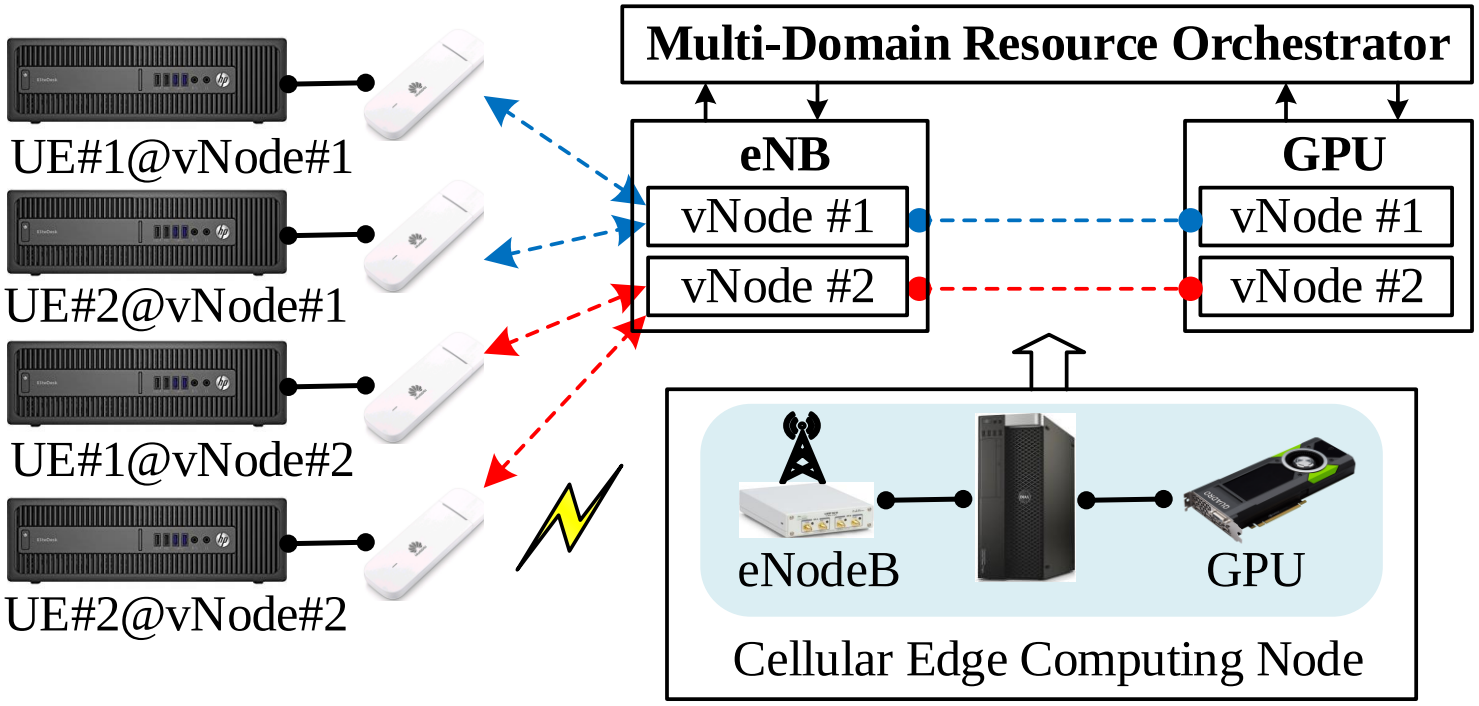


1. Slice tenant requests a network slice
2. Admission control of slice requests
3. Orchestrator allocates the multiple domain resources (virtual) for all admitted network slices
4. Each network slice allocates resources (virtual) to its users
5. The virtual resource allocations of all the users are informed to hypervisor
6. The hypervisor maps the virtual resources to physical resources to maximize the efficiency of physical resources

Qiang Liu and Tao Han, "VirtualEdge: Multi-Domain Resource Orchestration and Virtualization in Cellular Edge Computing", IEEE International Conference on Distributed Computing Systems (ICDCS) 2019.

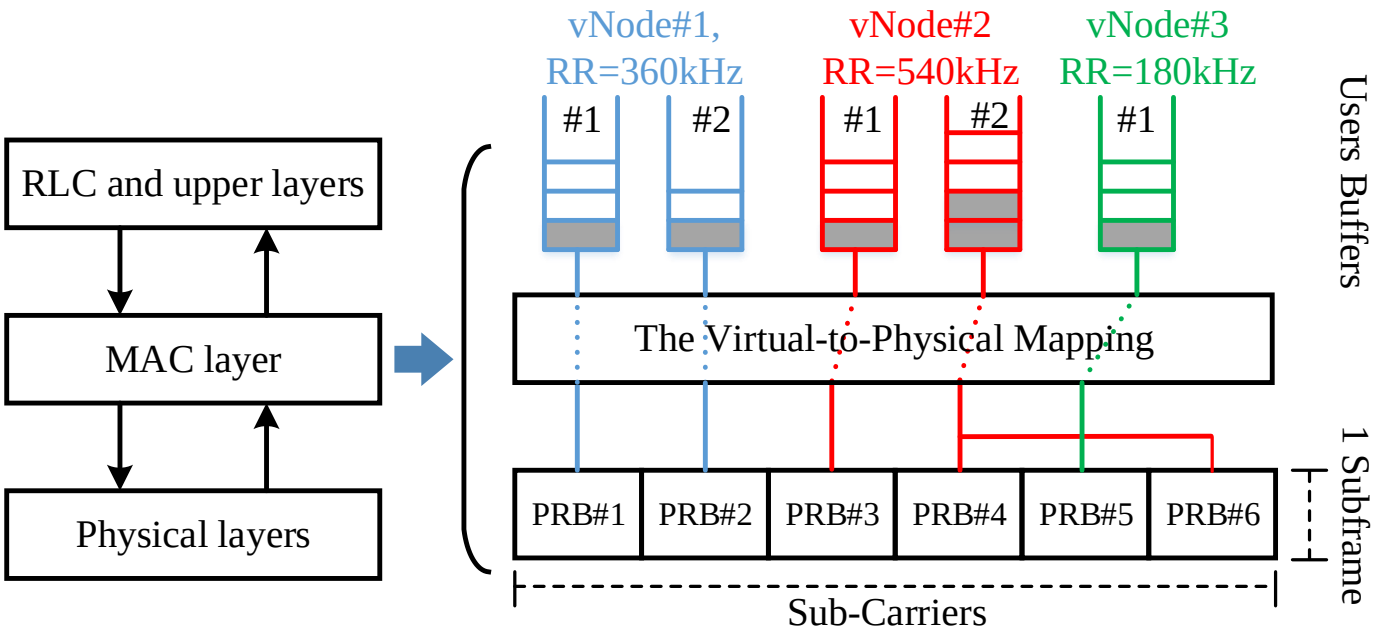
System Implementation

The system is developed and implemented based on the OpenAirInterface (OAI) LTE and CUDA GPU computing platforms



Radio Resource Hypervisor

- Managing the MAC layer user scheduling and resource allocation (physical resource blocks (PRBs) in LTE network).



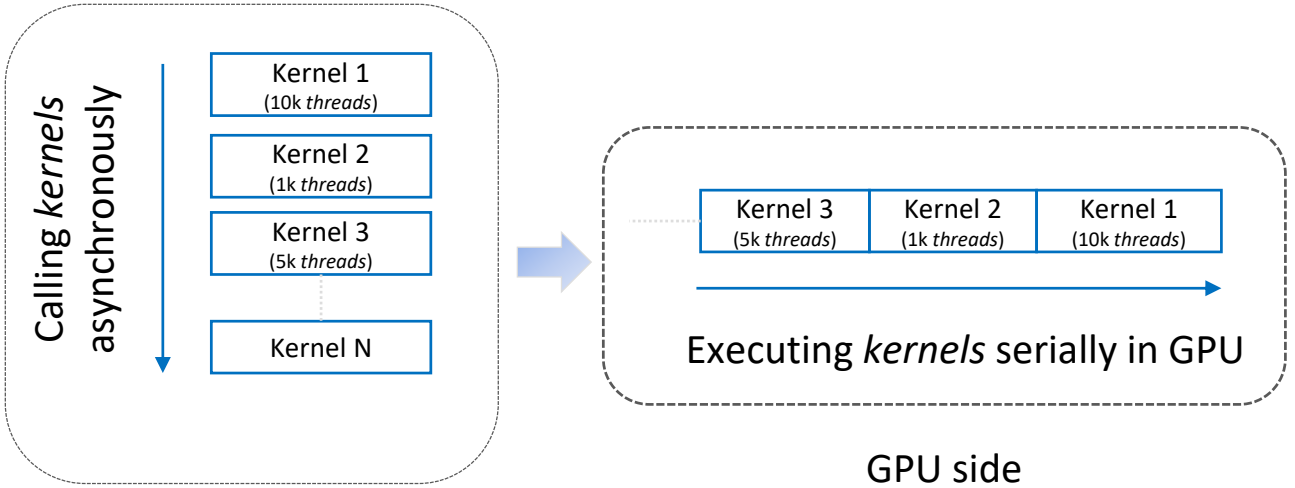
Computing Resource Hypervisor

- Managing the dispatch of kernel functions (Token-based)

Kernel function in CUDA programing:

```
MyKernel<<<BLKNUM,THDNUM>>>(parameters);
```

name *threads* parameters



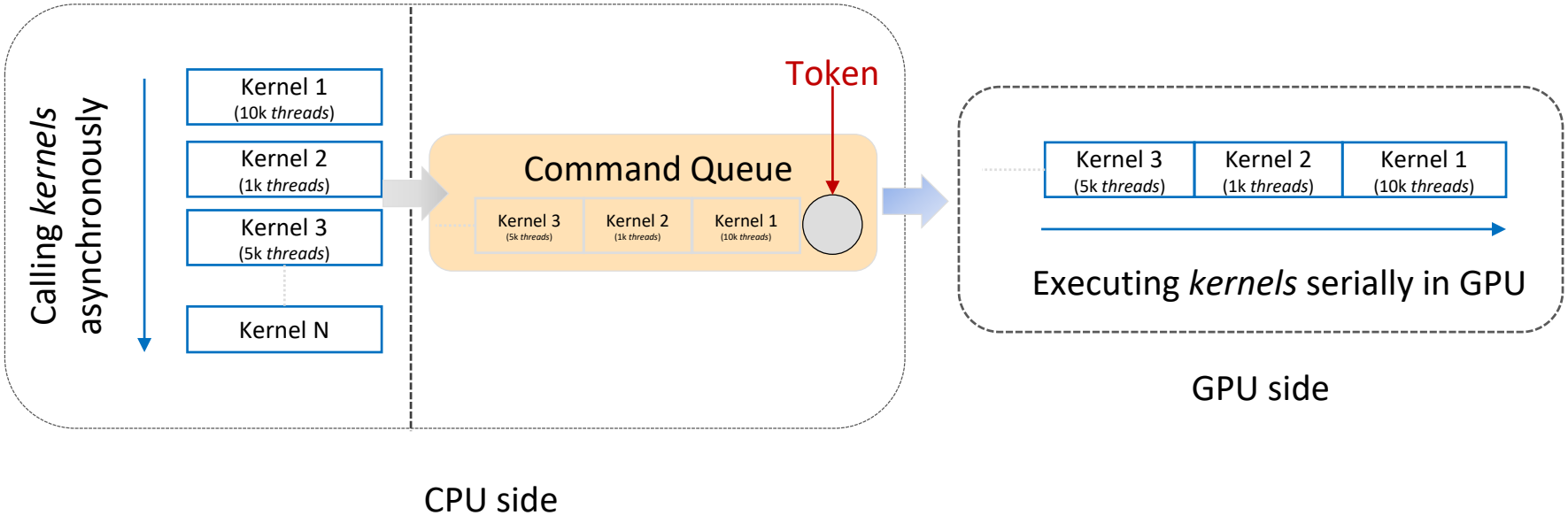
Resource Hypervisor: Computing

- Methodology: Managing the dispatch of kernel functions (Token-based)

Kernel function in CUDA programing:

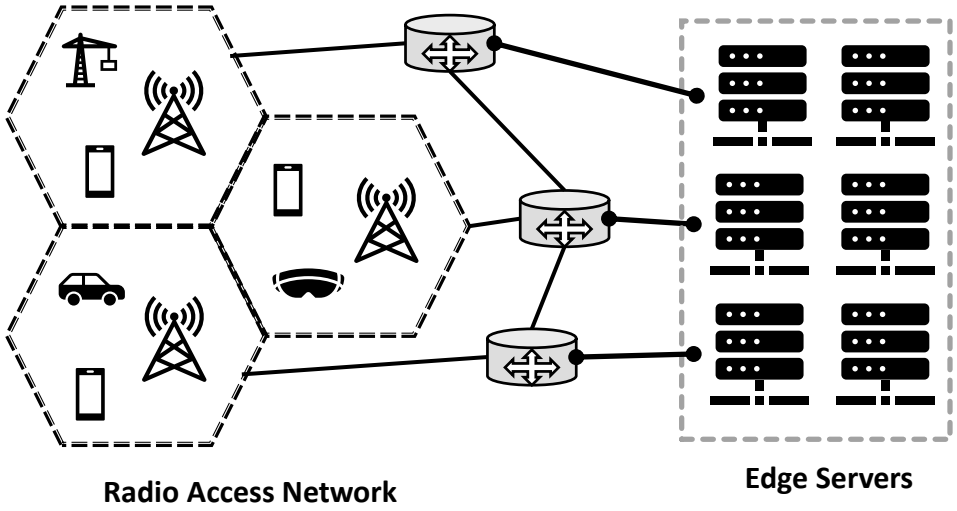
```
MyKernel<<<BLKNUM,THDNUM>>>(parameters);
```

name *threads* parameters



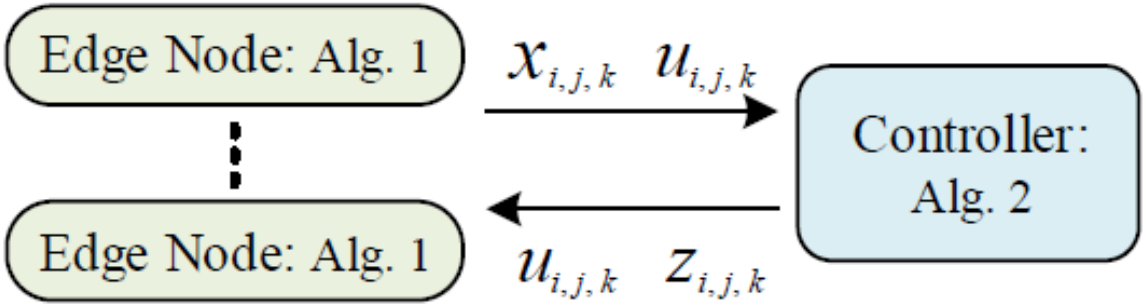
Distributed Resource Orchestration

- Considering multiple eNodeBs and computing servers



Qiang Liu and Tao Han, “DIRECT: Distributed Cross-Domain Resource Orchestration in Cellular Edge Computing”, ACM International Symposium on Mobile Ad Hoc Networking and Computing(MOBIHOC) 2019.

Algorithm Overview



- Controller side:
  - Updating the dual variables and optimize the auxiliary variable Z (convex problem)
- Node side:
  - Optimize the resource allocation X

System Overview

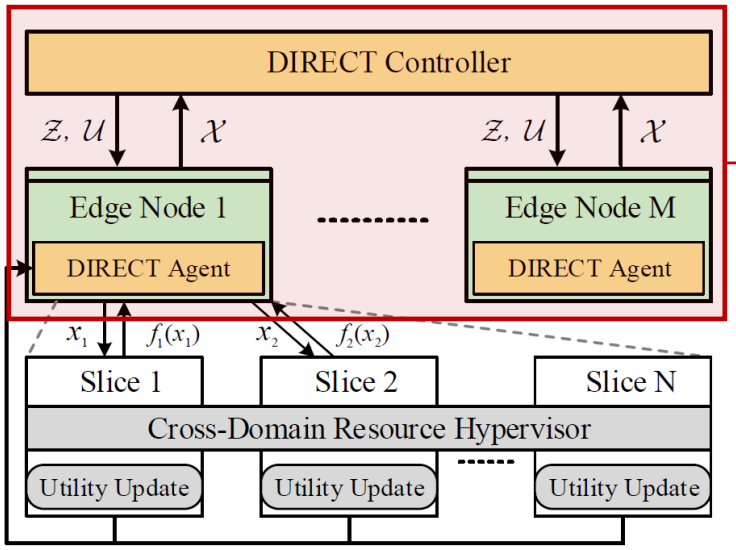


Figure 3: The design of DIRECT protocol.

- Slice Orchestrator: Dynamically orchestrate virtual network resources to slices across the network

- DIRECT controller: Coordinate the resource allocations to slices across edge nodes (control-side algorithm)
- DIRECT agents in edge nodes: Allocate resources to slices using a learning-based algorithm (edge-side algorithm)

System Overview

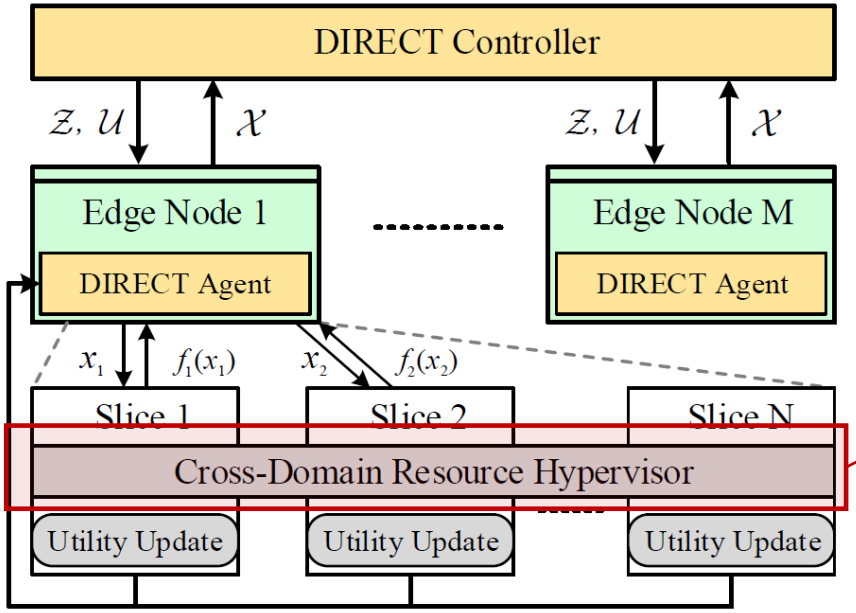


Figure 3: The design of DIRECT protocol.

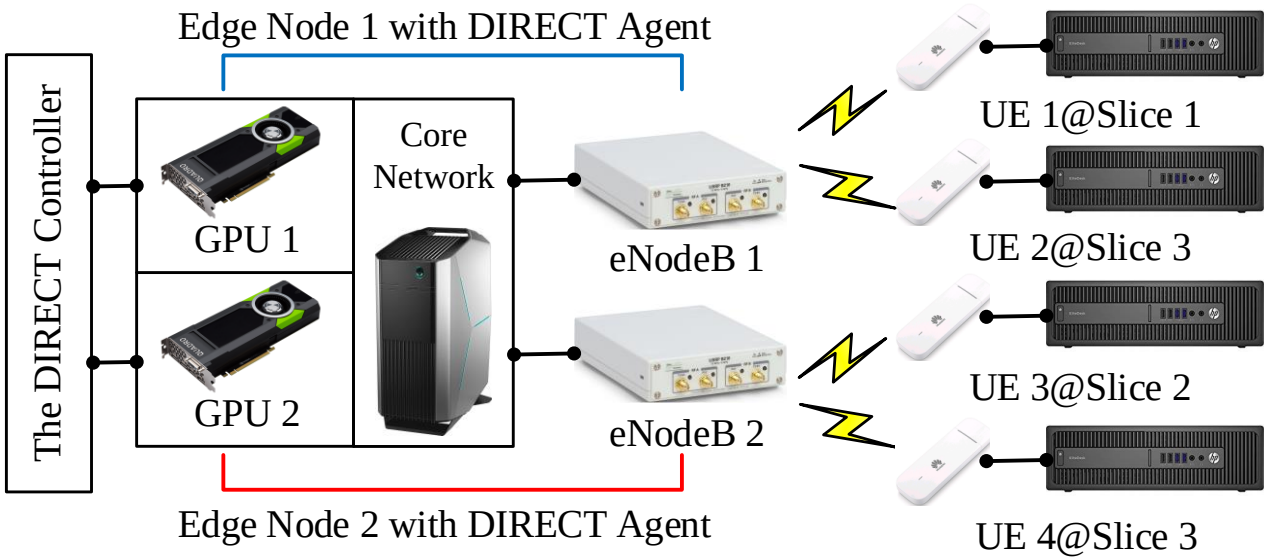
➢ Resource Hypervisor: map virtual resource to physical resources

Radio Resource Hypervisor

Computing Resource Hypervisor

System Implementation

- Hardware Details:
  - OpenAirInterface (OAI): 2x USRP B210 SDR boards, 2x eNodeB computers, 1x Core network
  - CUDA GPU computing platform: 2x NVIDIA GTX 1080Ti, CUDA 8.0
  - Mobile users: 4x Huawei dongle E2273, 4x Linux computers

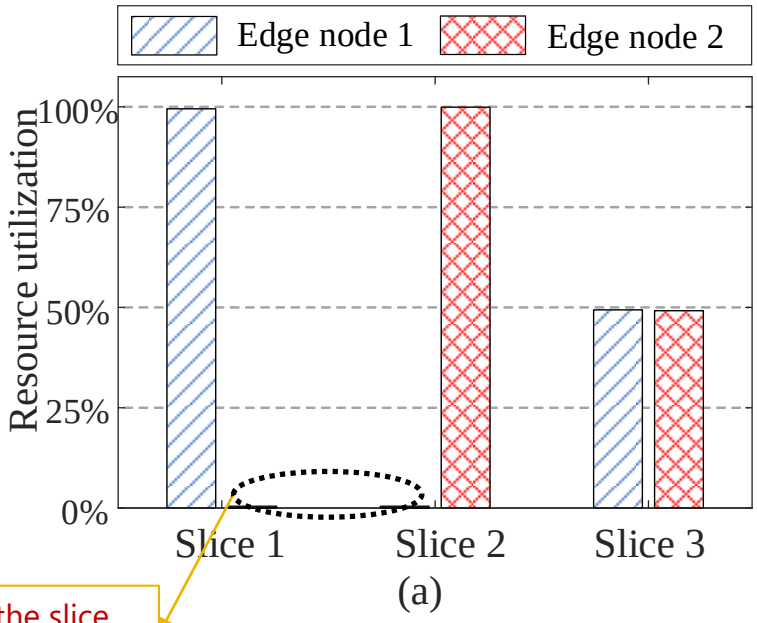


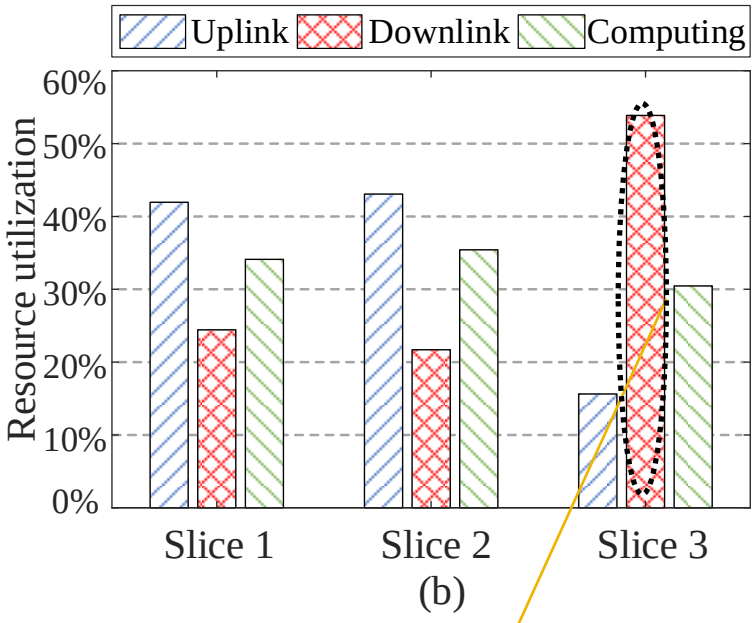
Experiment Results

Table 1: User Association

	Slice 1	Slice 2	Slice 3
Edge Node 1	1	0	1
Edge Node 2	0	1	1
application	MAR	MAR	VAS

- DIRECT is aware of the traffic load
- DIRECT learns the needs of multi-domain resources of slices

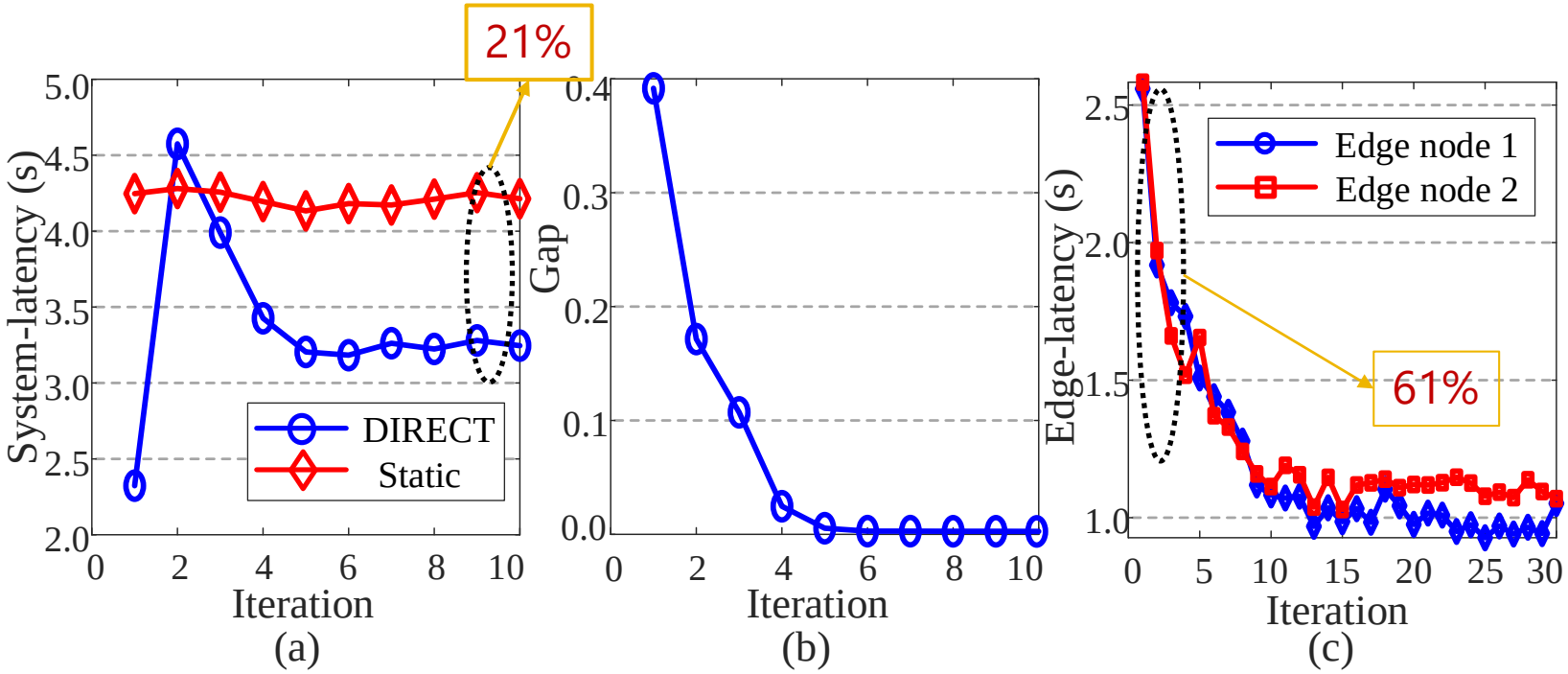




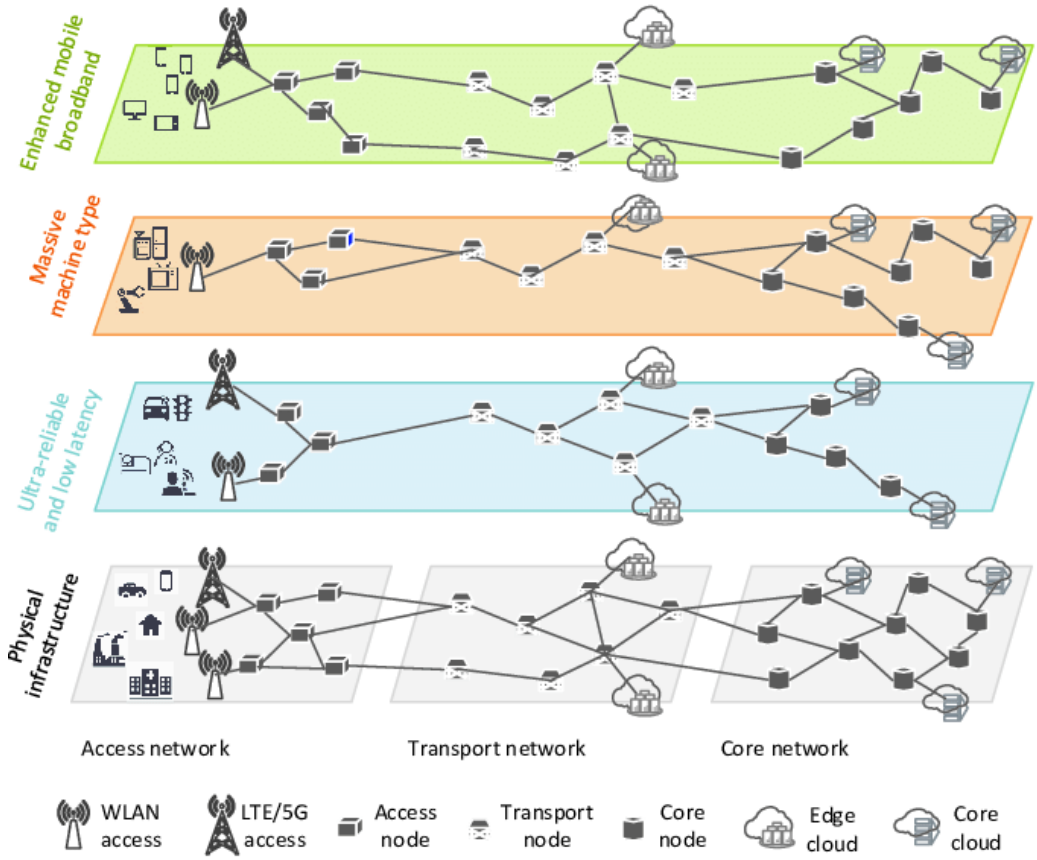
Learn the slice traffic on edges

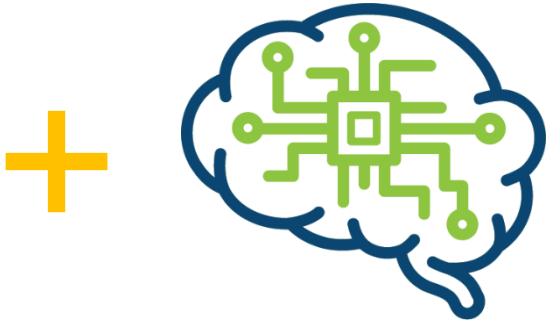
Learn the resource demand of application

Experiment Results



- DIRECT converges in a few iterations.
- DIRECT reduces about 21% system latency as compared to Static.
- DIRECT agents can learn the optimal resource allocations to network slices.





AI

Tao Han

The University of North Carolina at Charlotte

Tel #:704-687-8406, Email: Tao.Han@uncc.edu