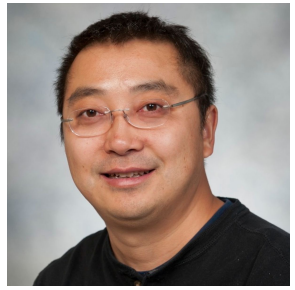
Agora: real-time massive MIMO baseband processing in software

Jian Ding[*], Rahman Doost-Mohammady[✣], Anuj Kalia[★], Lin Zhong[*]

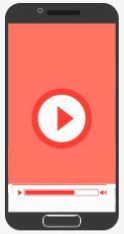
[*]Yale University [✣]Rice University [★]Microsoft



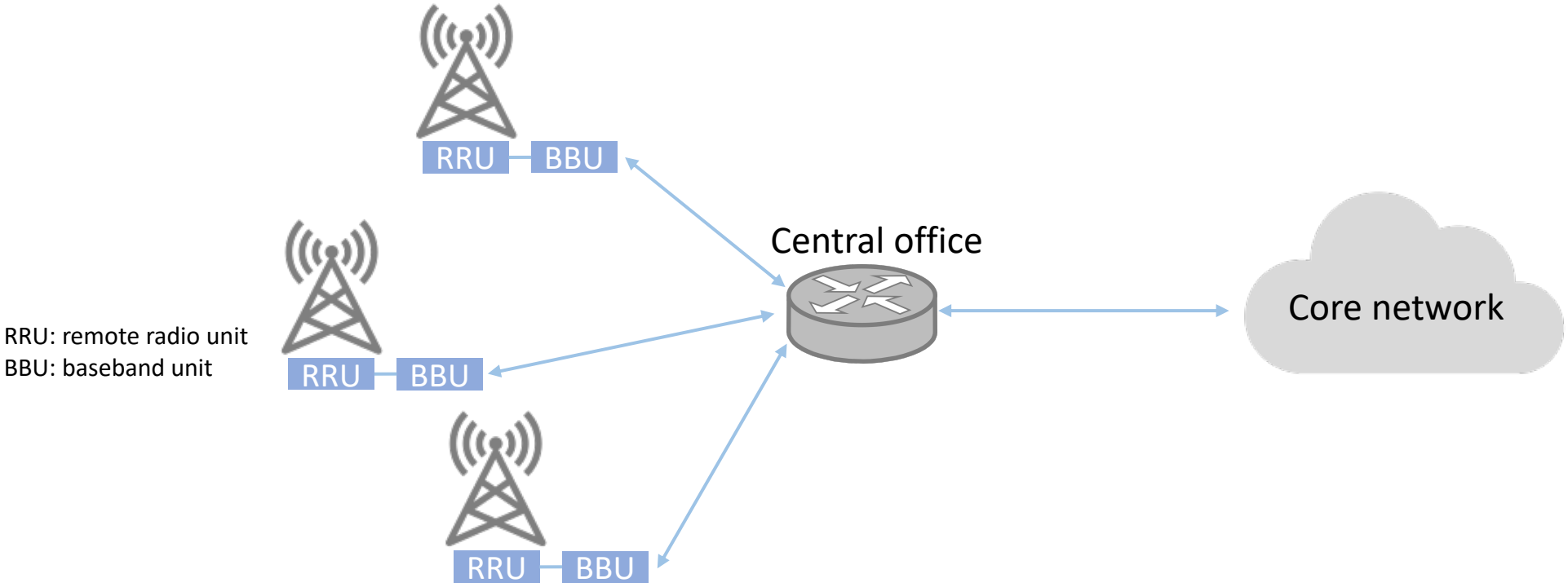
CONEXT 2020
2020.12.03

Moile network evolution: A decade per G

1980s	1990s	2000s	2010s	2020 onwards
1G	**2G**	**3G**	**4G**	**5G**
Voice	Voice \ Text	Voice \ Text \ Internet access	Voice \ Text Internet \ Video	Smart Home
				
s p e e d s				
2.4kbps	64kbps	2mbps	100mbps	1gbps
				

Why: dedicated, specialized, distributed equipment

Hard to program & replace, expensive (10^9)

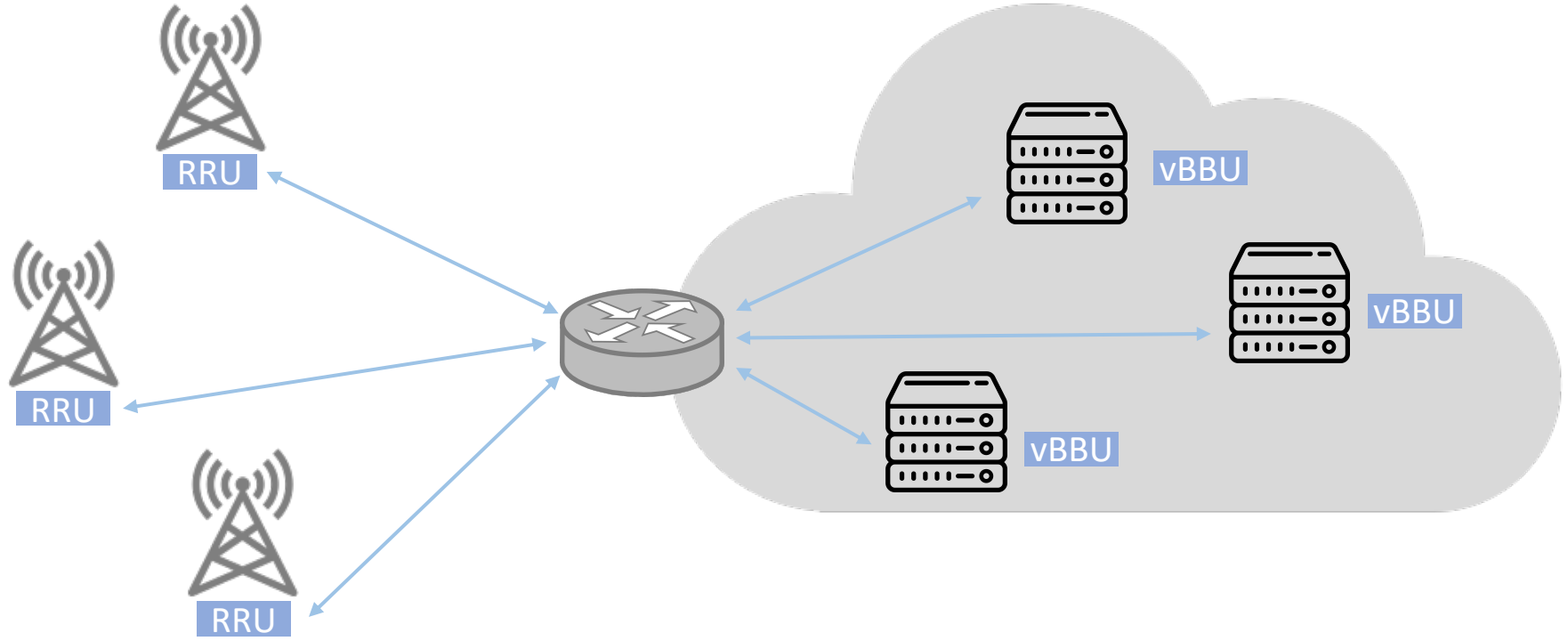


Traditional, distributed radio access network (RAN)

Solution: cloudify

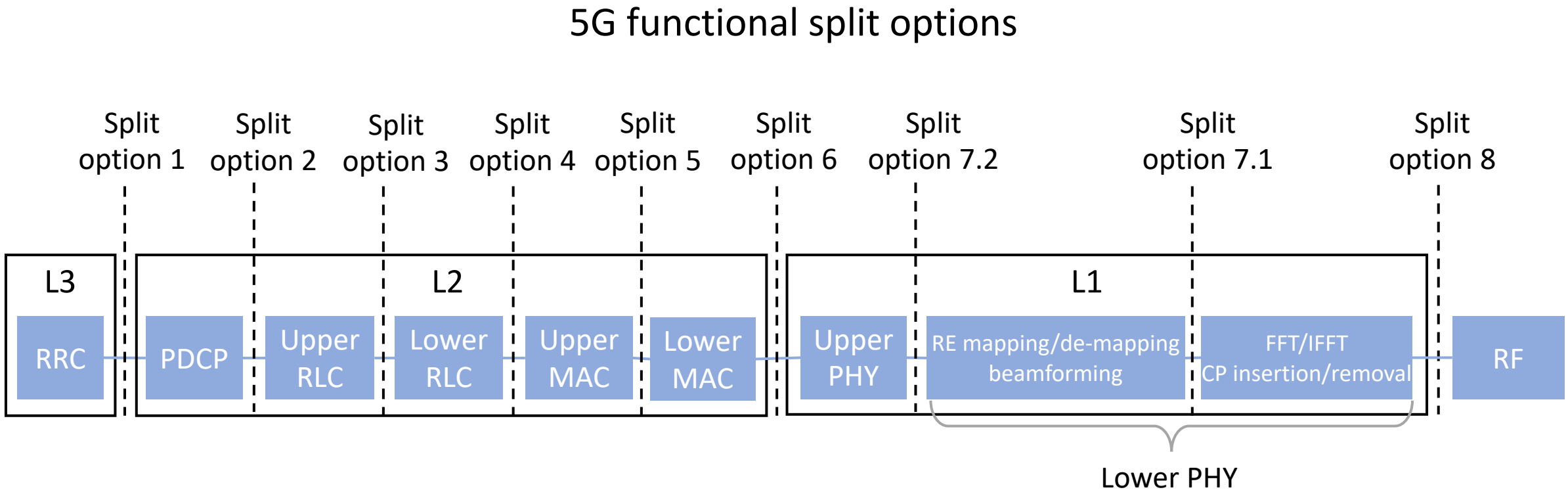
Only radio parts
stay on cell towers

Moving necessary computation to data centers

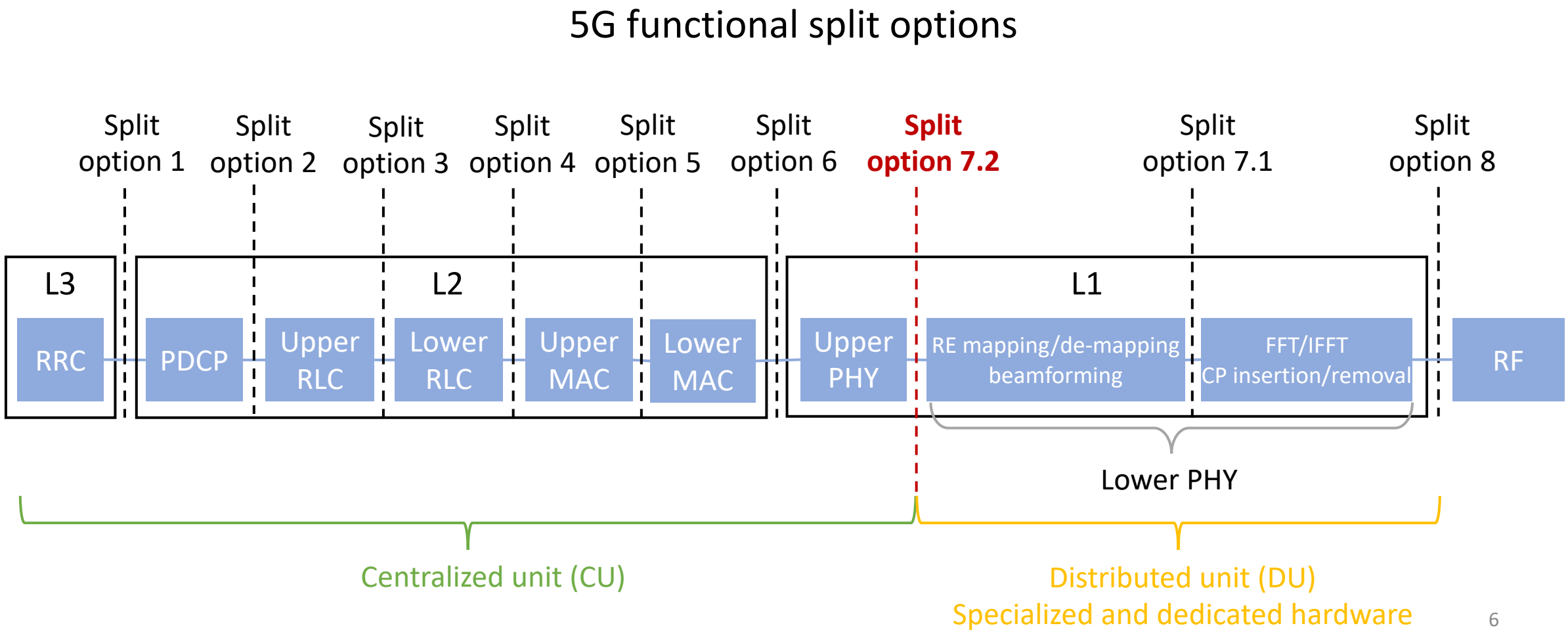


Virtual radio access network (vRAN) architecture

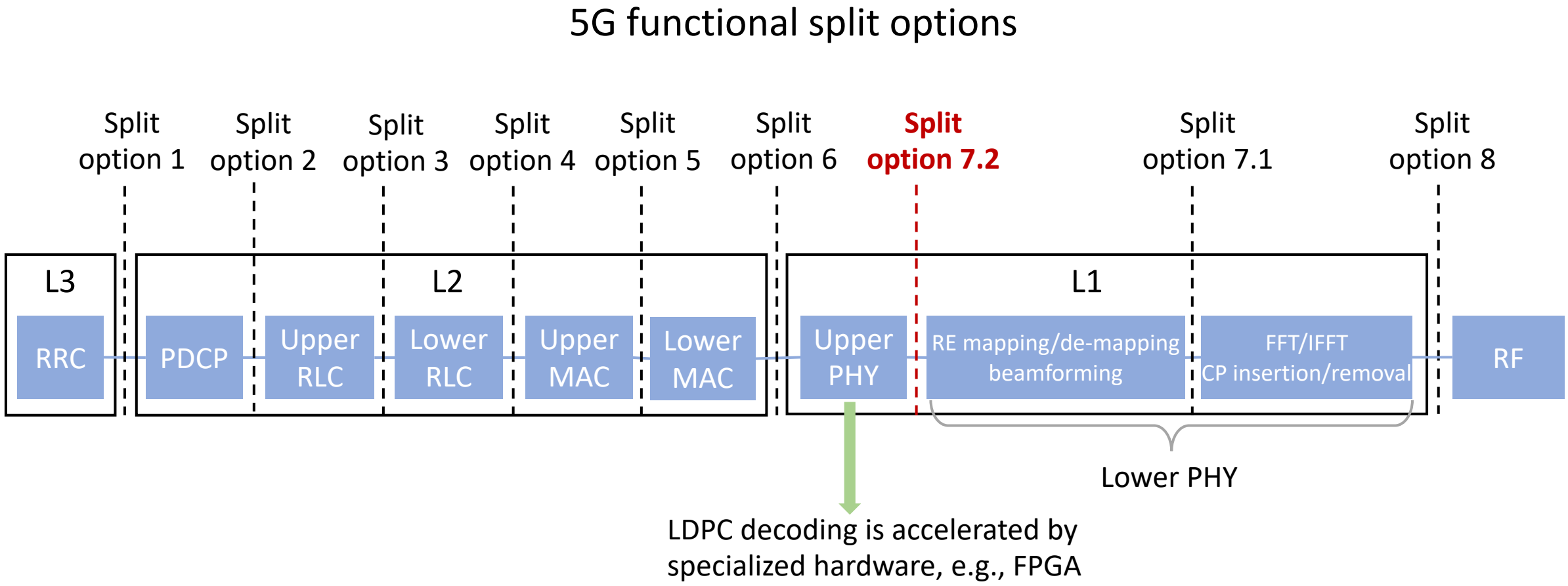
Existing vRAN solutions: most intensive computations (lower PHY) remain distributed



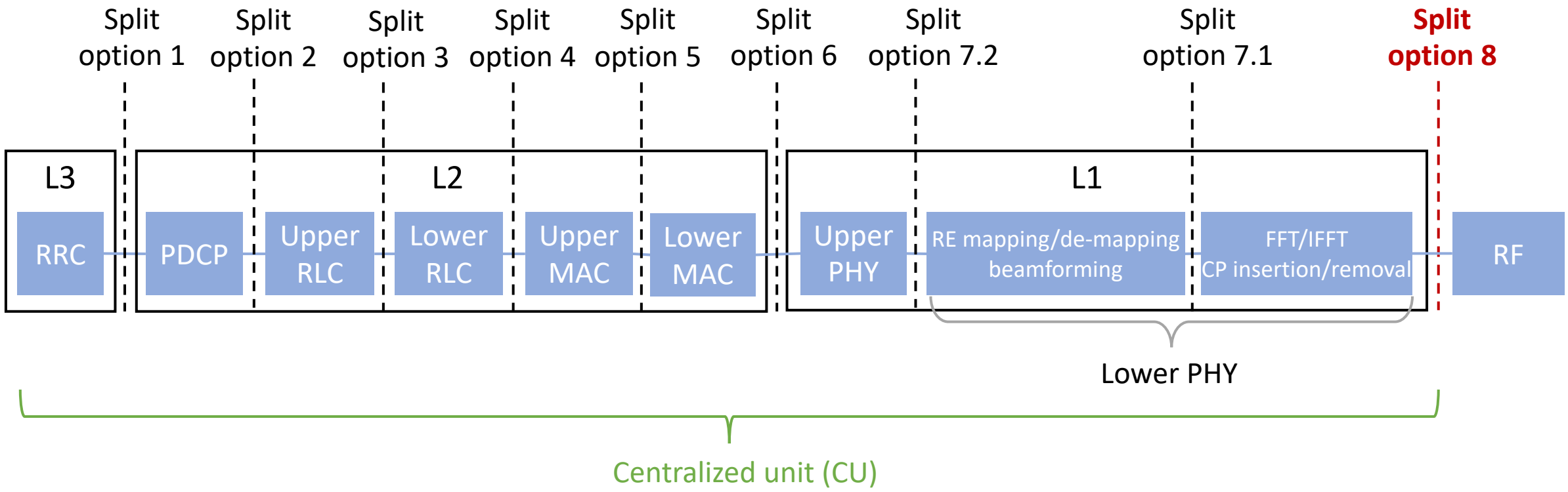
Existing vRAN solutions: most intensive computations (lower PHY) remain distributed



Existing vRAN solutions: intensive computations in CU rely on hardware acceleration



Can lower PHY also be centralized?

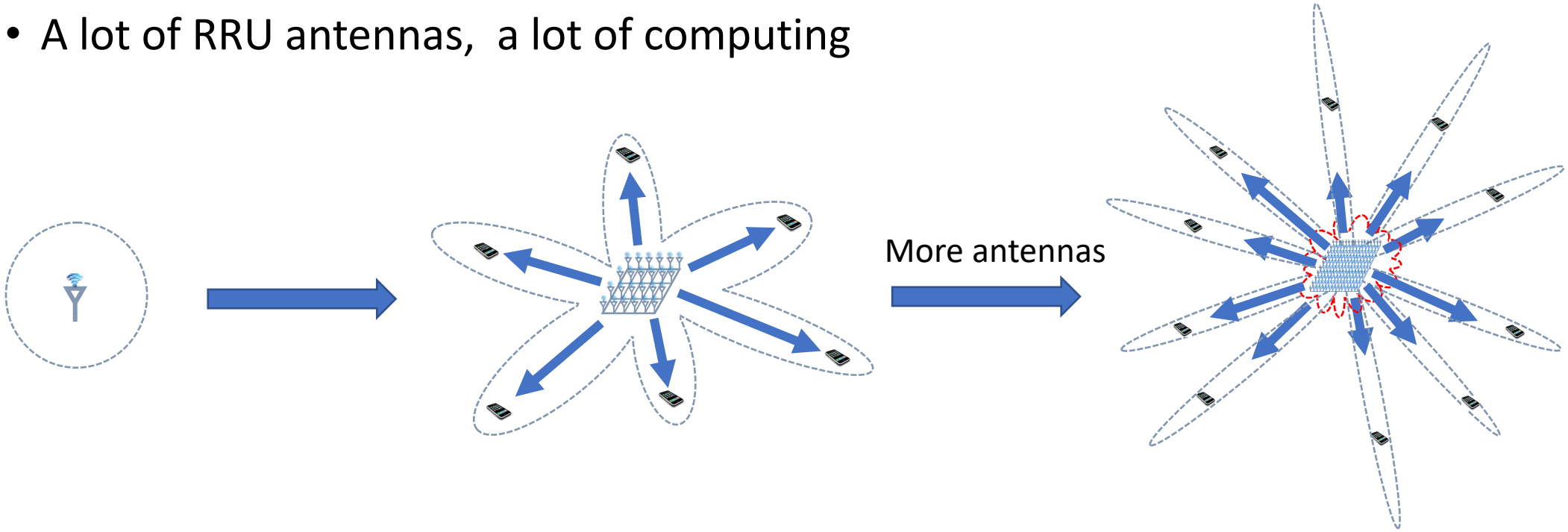


Challenge: high performance requirements of 5G NR

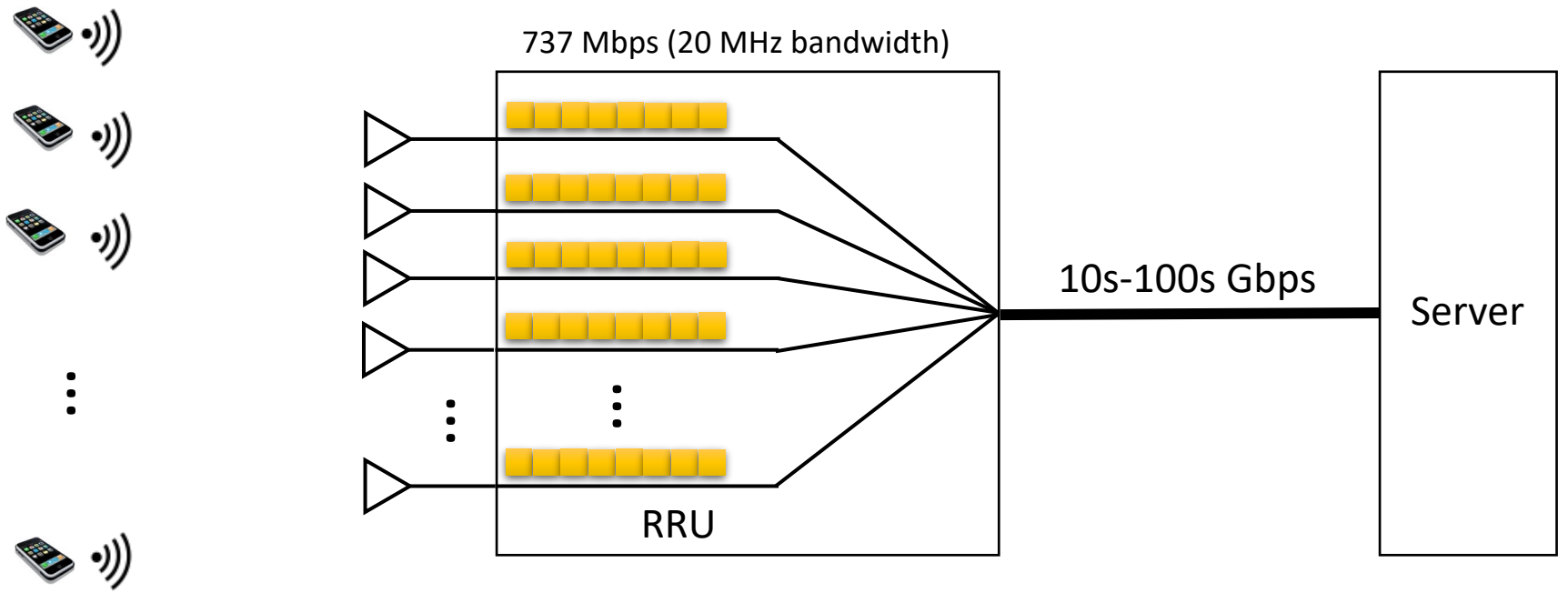
- Low end-to-end latency
  - < 1 ms for URLLC, < 4 ms for eMBB
- High data rate
  - > 3 Gbps with 100 MHz bandwidth (sub-6 GHz)

5G’s secret sauce: massive MIMO

- A lot of RRU antennas, a lot of computing

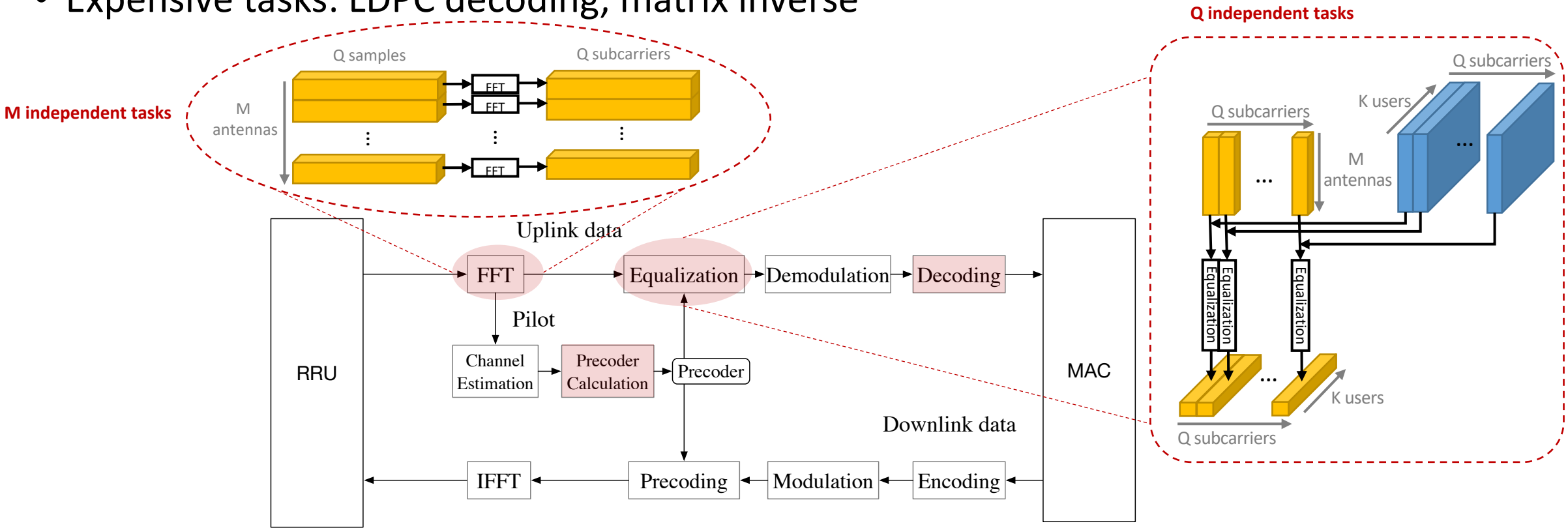


Challenge: massive MIMO is data-intensive



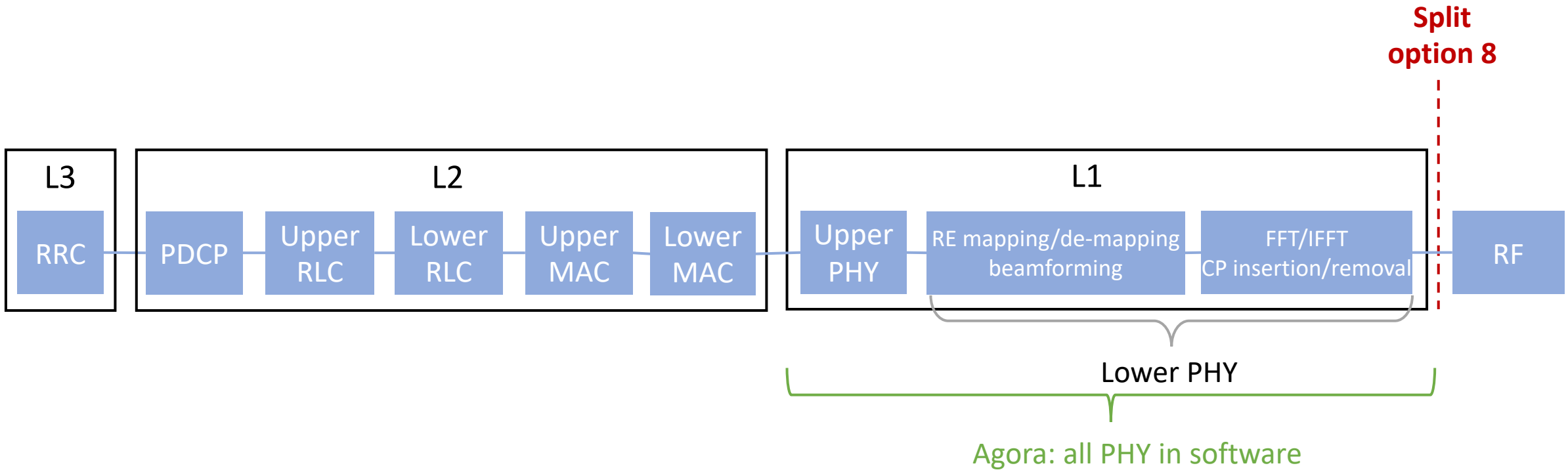
Challenge: massive MIMO baseband processing is compute-intensive

- Each **block** consists of many identical, independent **tasks**
- Expensive tasks: LDPC decoding, matrix inverse



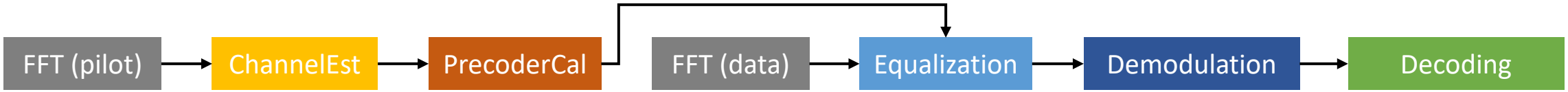
Agora: real-time massive MIMO baseband processing on a many-core server

- Split option 8 is feasible for 64 RRU antennas and 16 UEs using only 26 CPU cores
- Achieves high data rate and low latency to meet 5G NR requirements



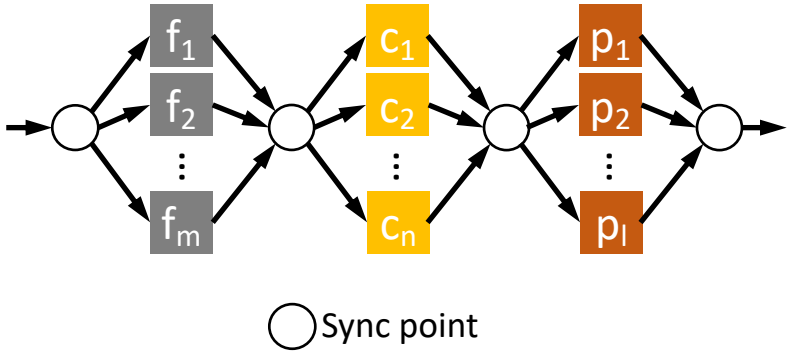
Key design principle: earliest frame first

- **Design choice:** always favor data parallelism over pipeline parallelism



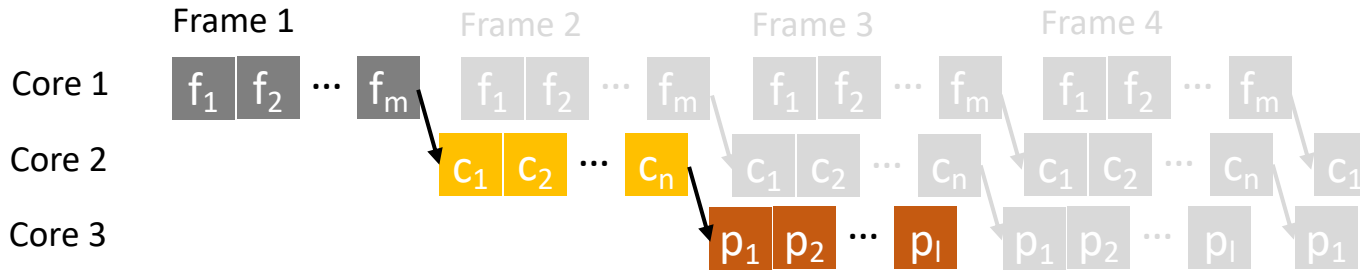
Data parallelism

Exists between identical, independent tasks



Pipeline parallelism

Exists between dependent blocks from different frames/symbols



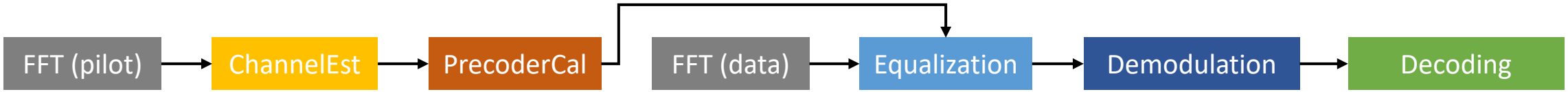
Each block takes longer time to process

f_i i-th independent FFT task

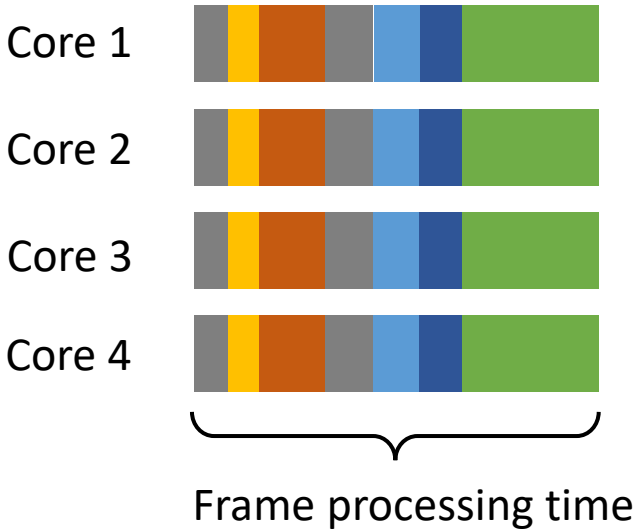
c_i i-th independent ChannelEst task

p_i i-th independent PrecoderCal task

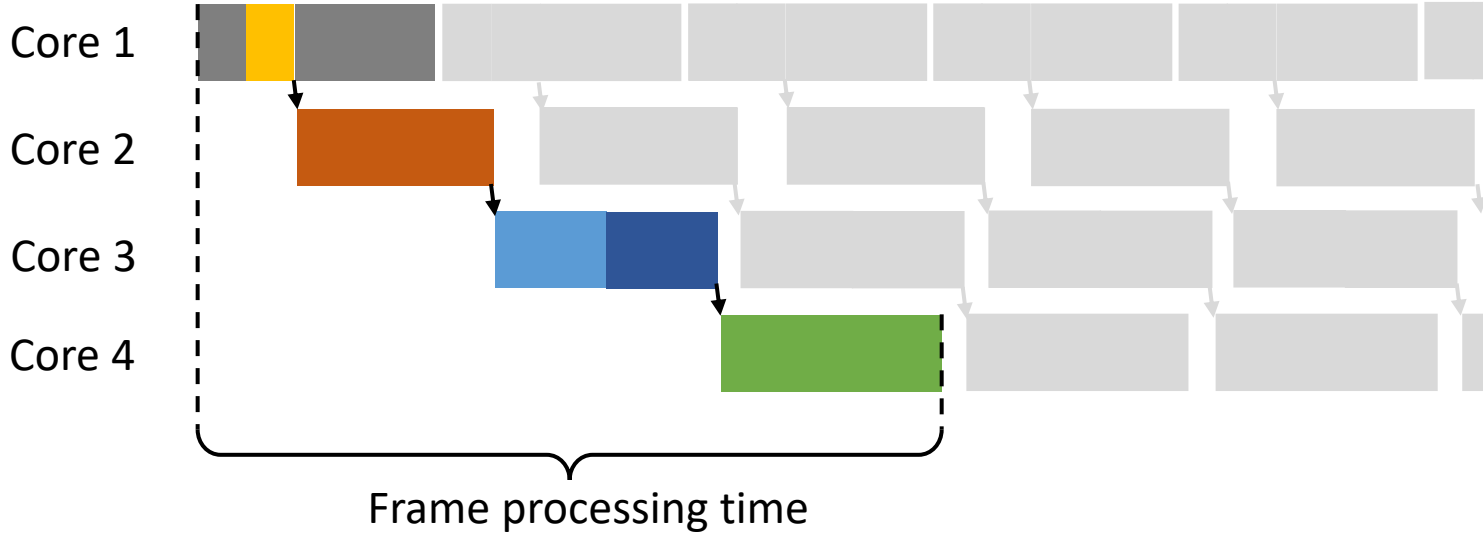
Data-parallel design gives shorter frame processing time



Data-parallel design



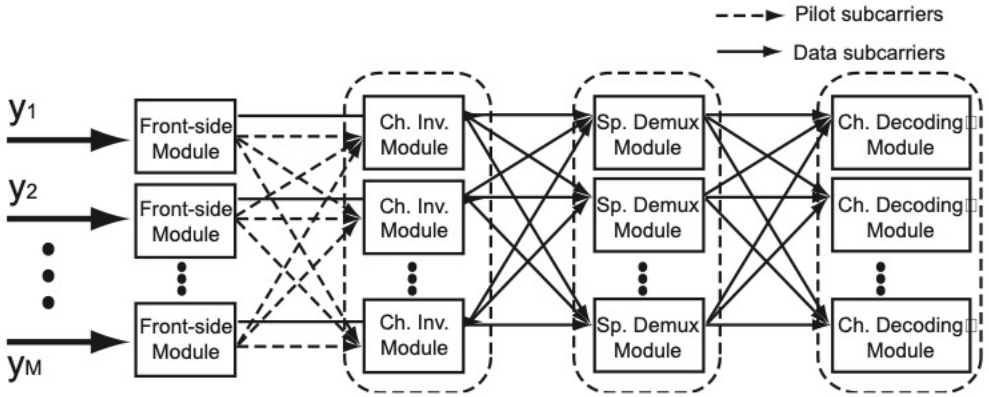
Pipeline-parallel design



Prior work favors pipeline parallelism

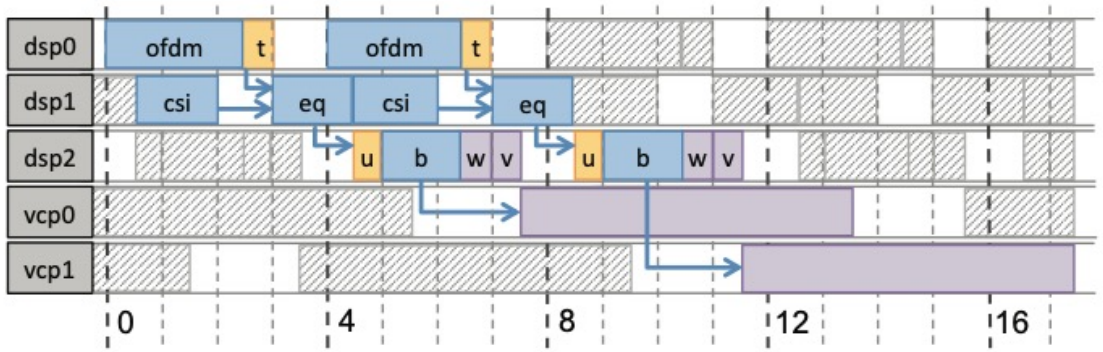
- Pipeline parallelism was necessary due to hardware architecture constraints

BigStation [SIGCOMM'13]



12x9 MIMO on distributed servers each with 4 cores

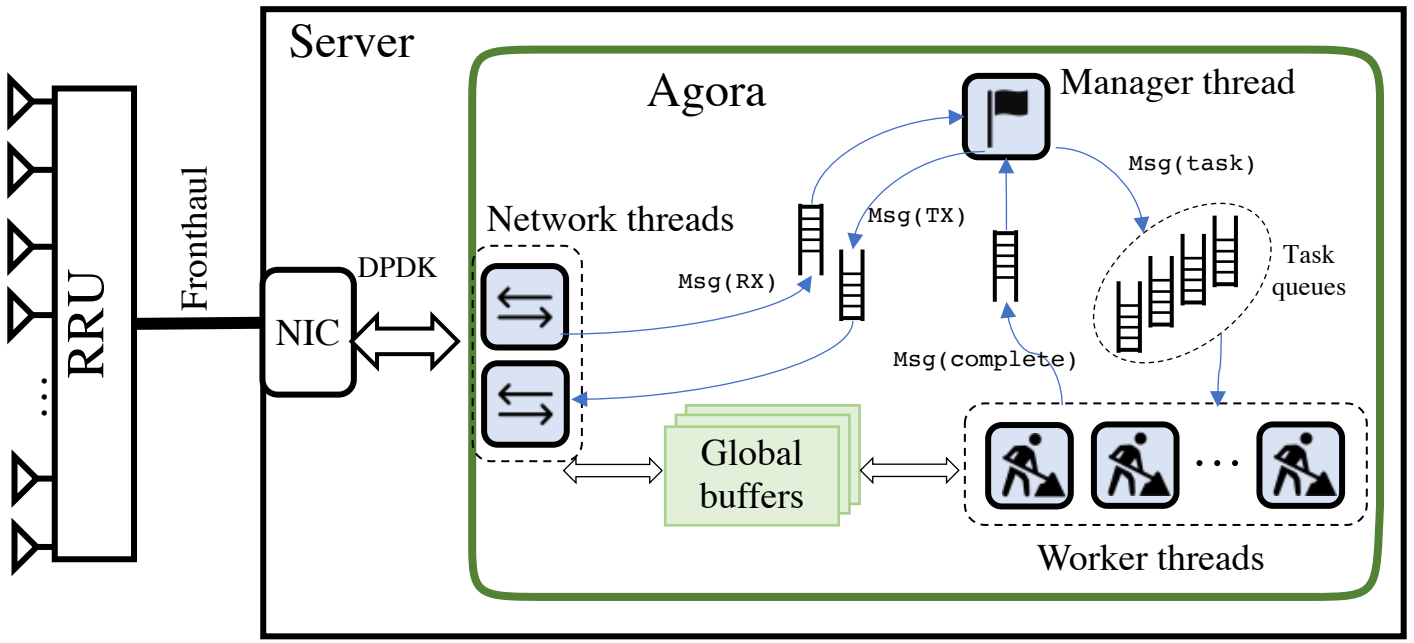
Atomix [NSDI'15]



802.11a Wi-Fi on DSP

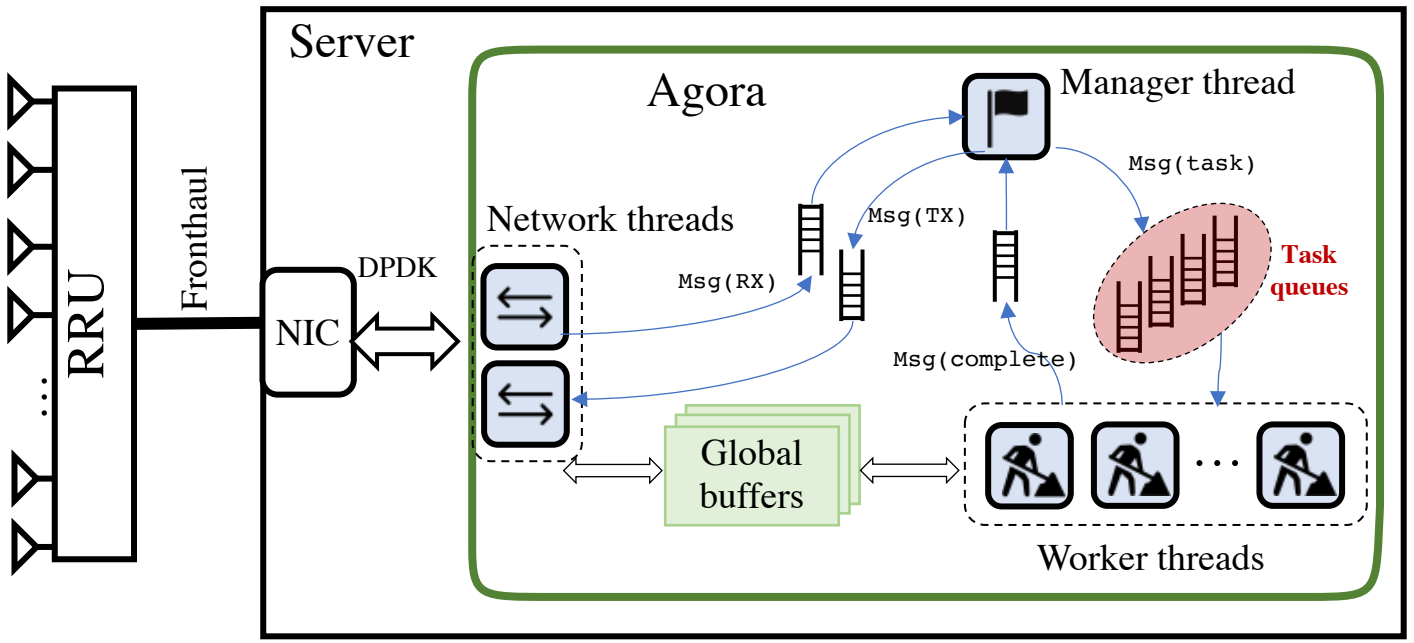
Challenge: achieve high CPU utilization while mapping data processing blocks to cores

- **Solution:** lock-free queue-based manager-worker threading model



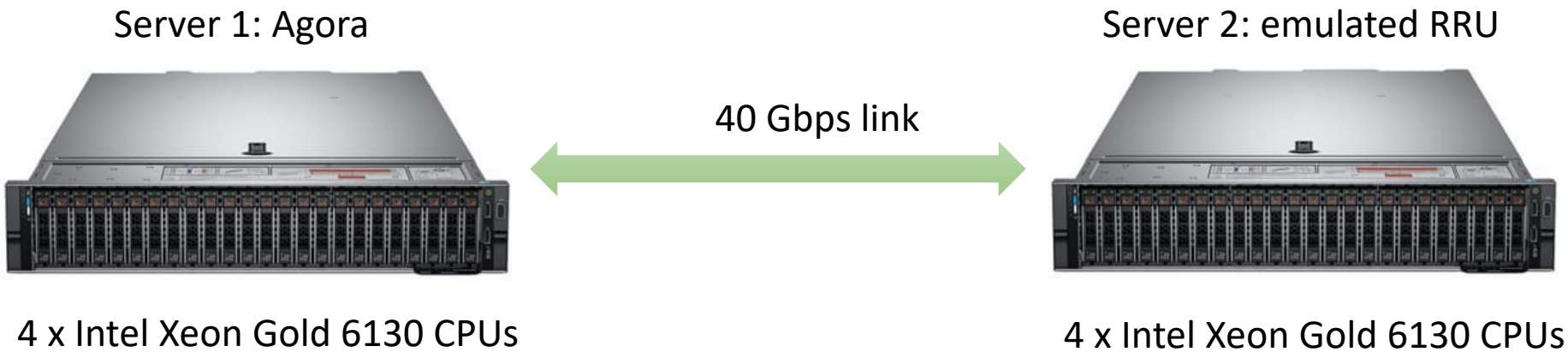
Apply earliest frame first principle in the threading model

- **Design choice:** enforce priority for the processing order of blocks
  - Use sperate task queues for different blocks; statically assign priorities to queues
  - Workers drain queues with higher priorities first



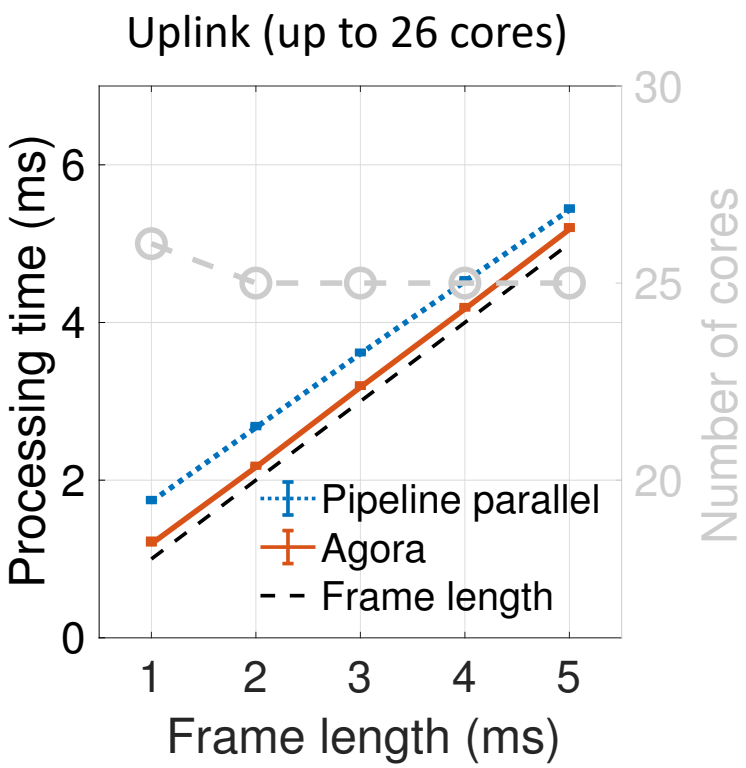
Experiment setup with emulated RRU

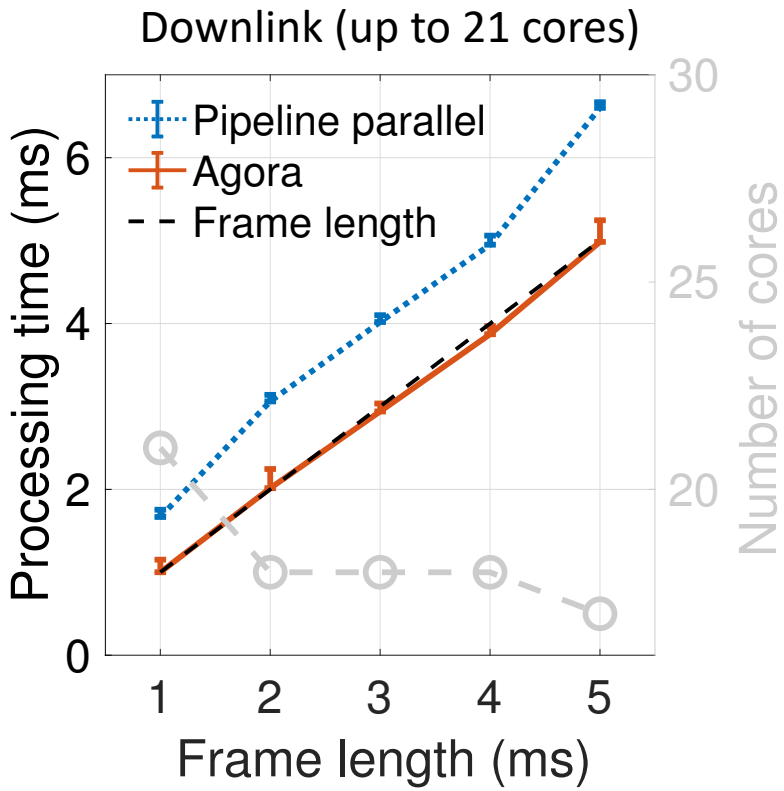
- Stress testing: 20 MHz bandwidth, 64-QAM, 1/3 LDPC code rate



Software-based massive MIMO can be real-time with low latency

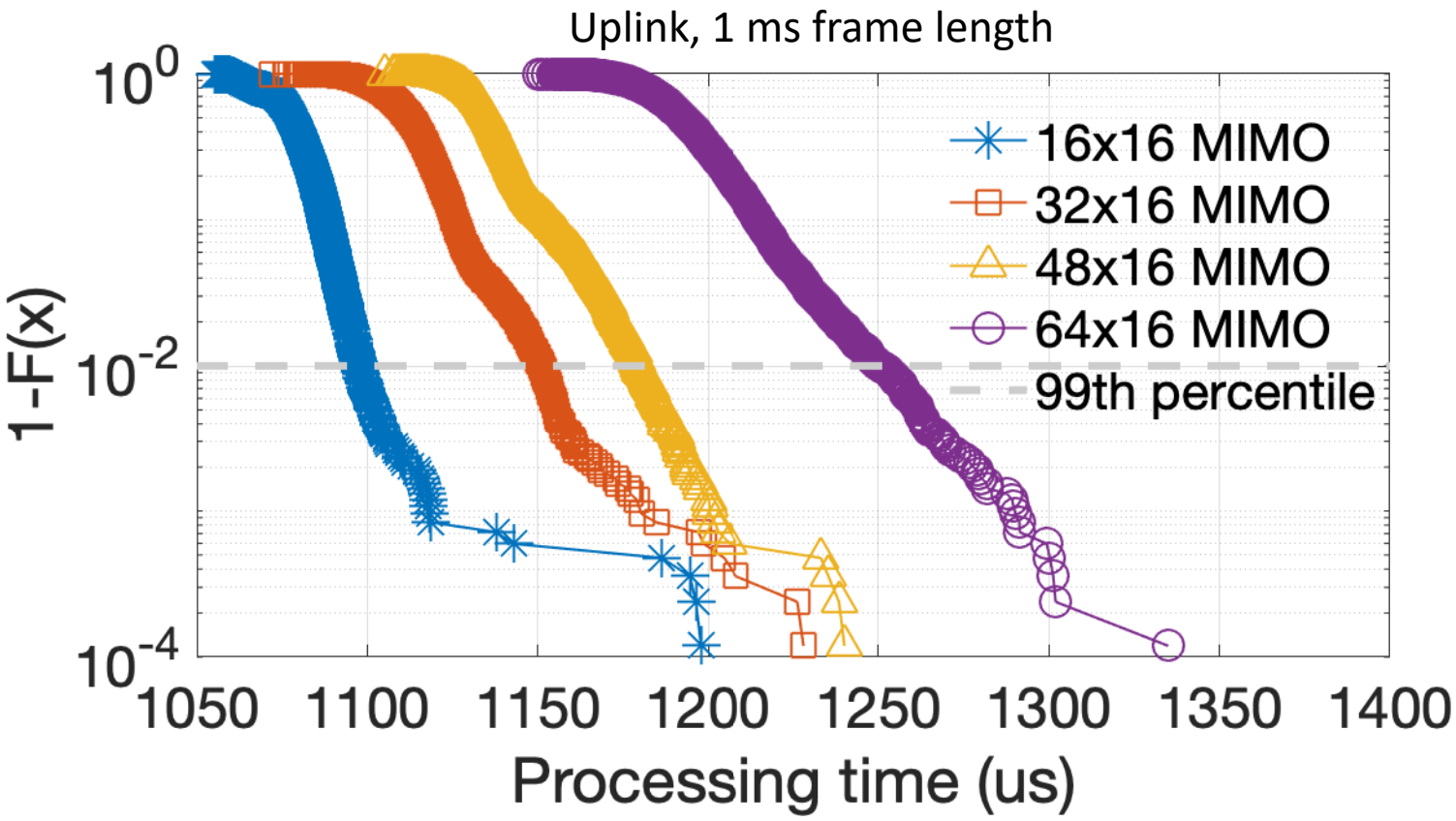
- Agora's frame processing latency of 64x16 MIMO is close to frame length





Agora achieves short tail latencies

- Meet 5G's latency requirement for eMBB, i.e., 4 ms



Server configurations are crucial for short tail latencies

Process priority of Agora	Median latency (ms)	Increase	99.9th latency (ms)	Increase
Real-time process	1.19	-	1.29	-
Normal process	1.16	0.98x	4.78	**3.71x**

Context switches are harmful

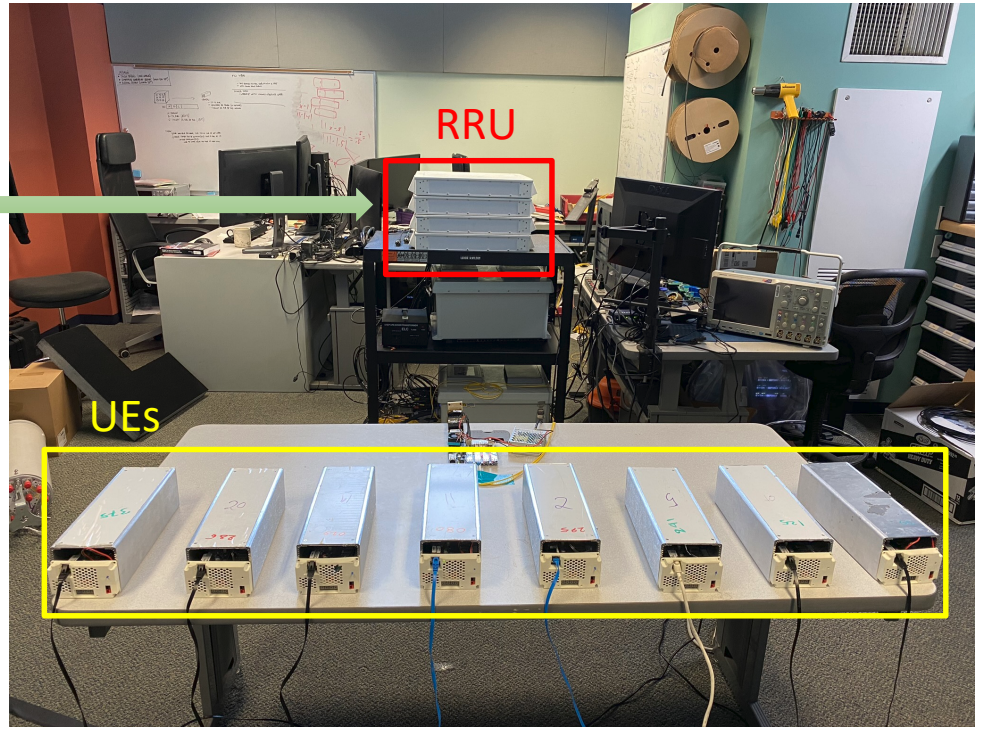
Experiment setup with real RRU and UEs

- 5 MHz bandwidth, 64-QAM, 1/3 LDPC code rate, indoor LoS

Server runs Agora

2 x Intel Xeon E5-2697 v4 CPUs

10 Gbps link

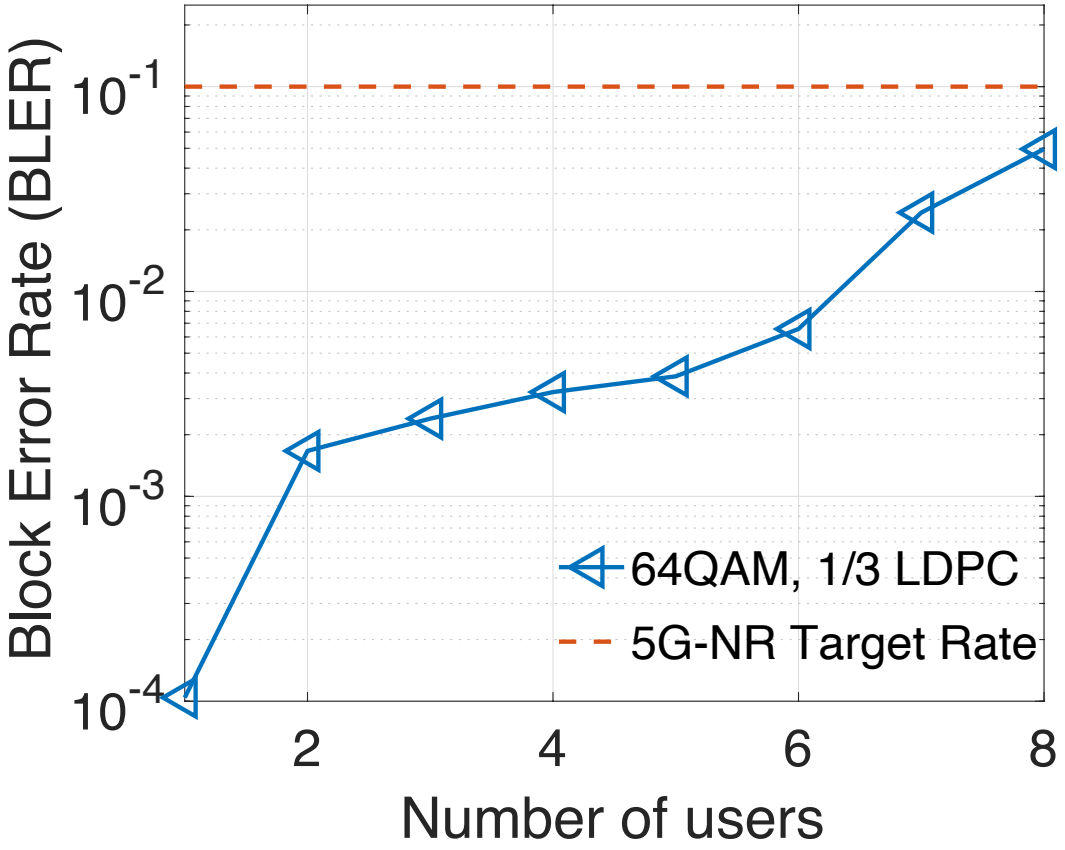


RRU and UEs are from Skylark Wireless

SKYLARK WIRELESS

Agora works with real RRU and UEs

- Block error rate meets 5G target with up to 64 RRU antennas and 8 UEs



Interesting things we learned

- ZF precoding is not really expensive as believed
  - 6 % of total CPU time for 64x16 MIMO
- LDPC decoding is a good candidate for accelerators
  - 37 % of total CPU time for 64x16 MIMO
- Inter-core communication overhead needs improvement for further scaling up massive MIMO
  - 34 % of total CPU time for 64x16 MIMO

Summary

- The first publicly known software realization of real-time massive MIMO baseband processing
- Up to 20 MHz, 64 x 16 MIMO with only 26 cores
- Meets low latency and high data rate required by 3GPP 5G NR eMBB
- Open-source at https://github.com/jianding17/Agora