Microsoft

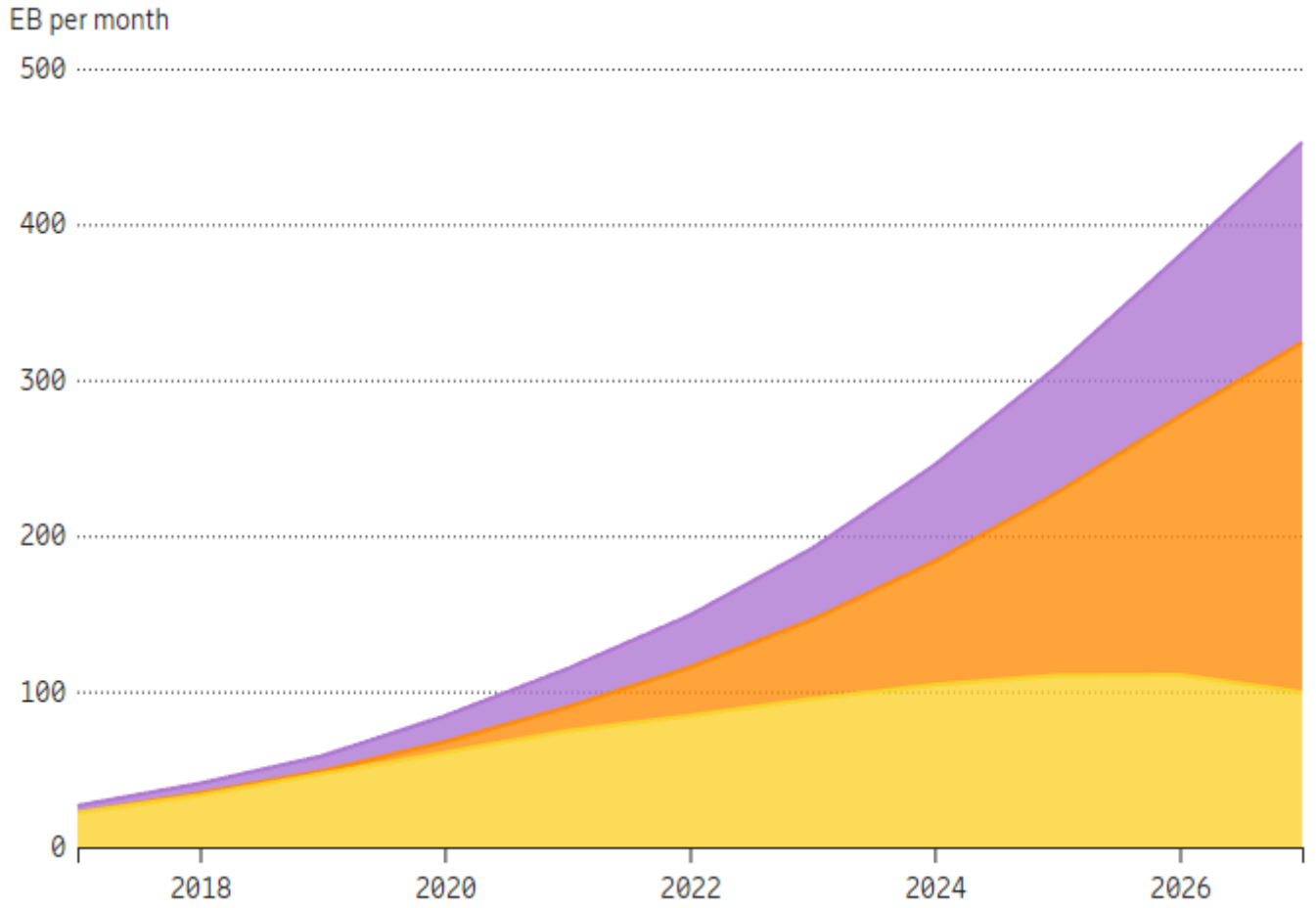
Scalable Distributed Massive MIMO Baseband Processing

Junzhi Gong (Harvard)

Anuj Kalia (Microsoft)

Minlan Yu (Harvard)

Increasing traffic rate in 5G

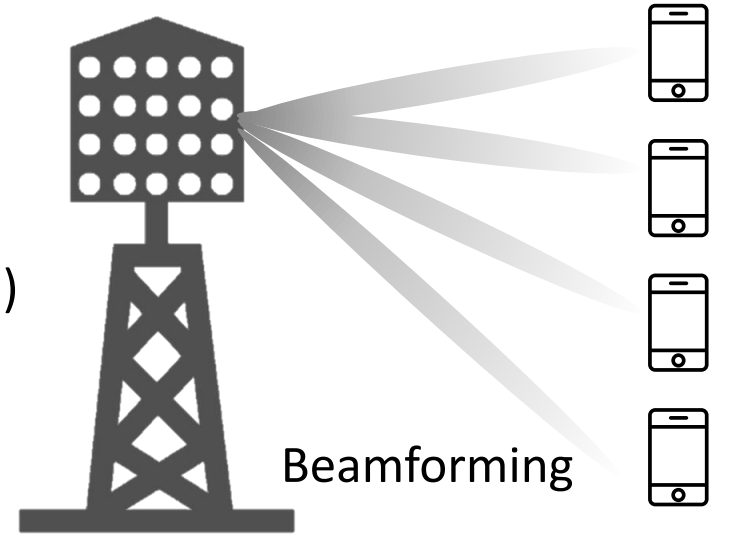


Increased demand on mobile traffic rate

Image credit to Ericsson

Key to higher data rate in 5G: massive MIMO

Radio unit (RU)

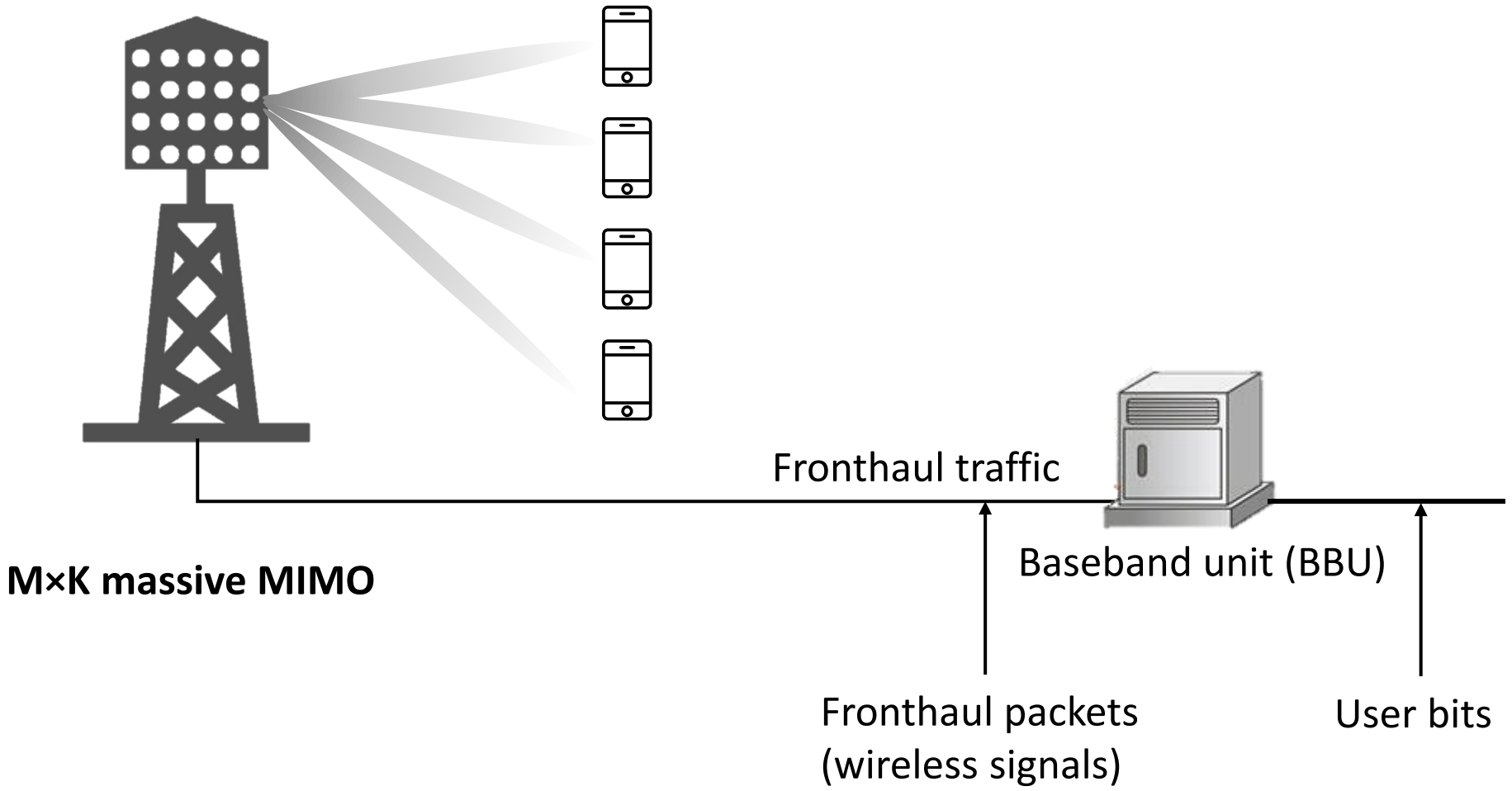


M×K massive MIMO

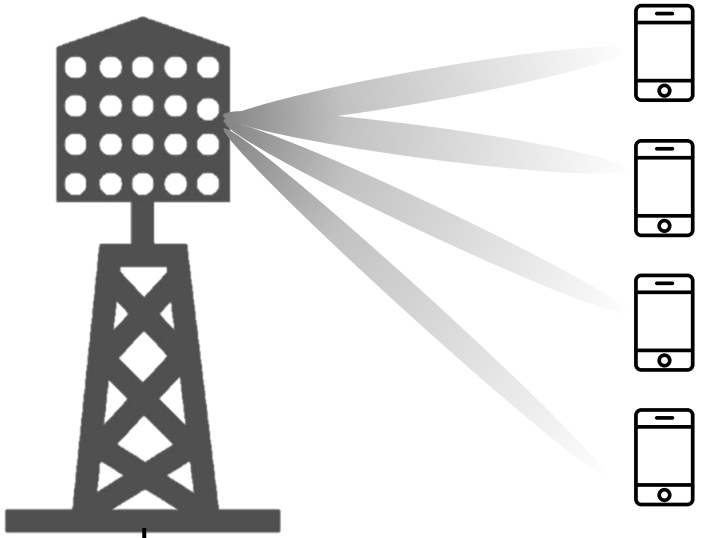
Massive MIMO: with many antennas, many users can send/recv data at the same time, at the same frequency

Beamforming: focuses radio signals directly at the users, to eliminate interference

Computation and wired communication challenges of massive MIMO



Computation and wired communication challenges of massive MIMO



M×K massive MIMO

Samsung Reveals How Its vRAN Technology Has Evolved To Underpin the Networks of the Future

on August 10, 2021 Audio Share

In recent years, leading operators around the world have selected Samsung as their partner in the journey towards next-generation networks. Samsung is taking 5G innovation to the next level with its industry-leading virtualized Radio Access Network (vRAN) solution that will enable mobile operators to bring greater flexibility, scalability, and resource efficiency in network management.

QCT, Radisys and Intel Deliver vRAN Solution for 5G

Nokia Cloud RAN – at the leading edge of 5G

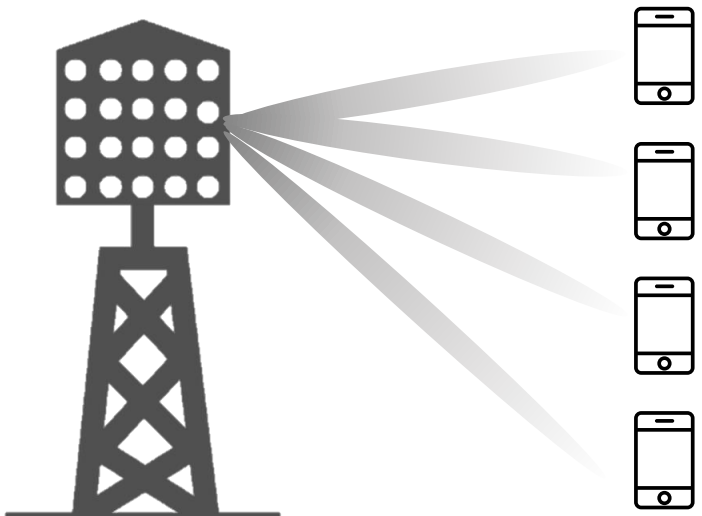
Nokia Cloud RAN enhances the potential of existing cloud infrastructure with real time 5G wireless connectivity. It fulfills the promise of a reliable connectivity application, fit for IT cloud environments. Nokia is actively collaborating with the industry ecosystem to further advance the development of 5G edge cloud solutions. We are a preferred partner of all three major hyperscalers: Amazon Web Services (AWS), Google, and Microsoft. Read more about our industry collaboration with these webscale companies here:

Fronthaul traffic

Commodity server

RAN virtualization movement

Computation and wired communication challenges of massive MIMO



Agora [CoNEXT 20]:

- Needs 28 cores (a full server) for 64x16 MIMO
- Can't scale to larger MIMO configurations (e.g., 128x32)

Single server is not enough

Large amount of computation with more antennas/users

Fronthaul traffic

Commodity server

M×K massive MIMO

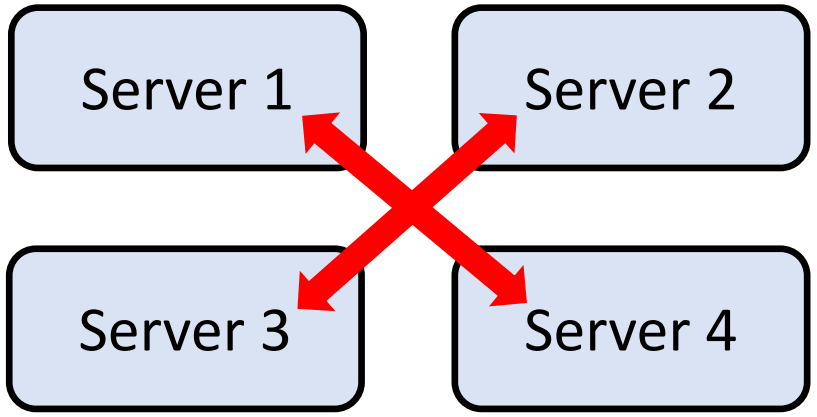
High fronthaul bandwidth with more antennas

Inter/intra-server communication limits scalability in prior massive MIMO systems

BigStation [SIGCOMM 10]

State-of-the-art distributed solution

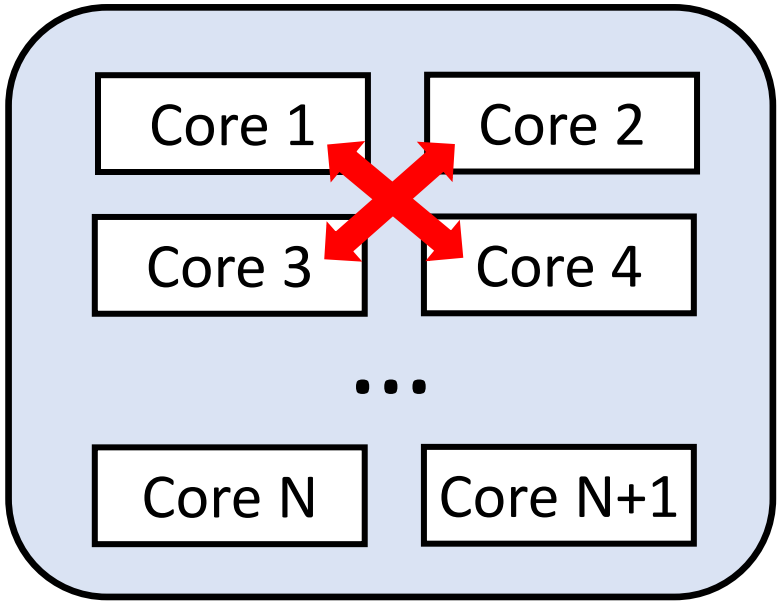
1 High inter-server communication



Agora [CoNEXT 20]

State-of-the-art single-server solution

2 High intra-server communication



Hydra: minimize inter and intra-server communication for scalability

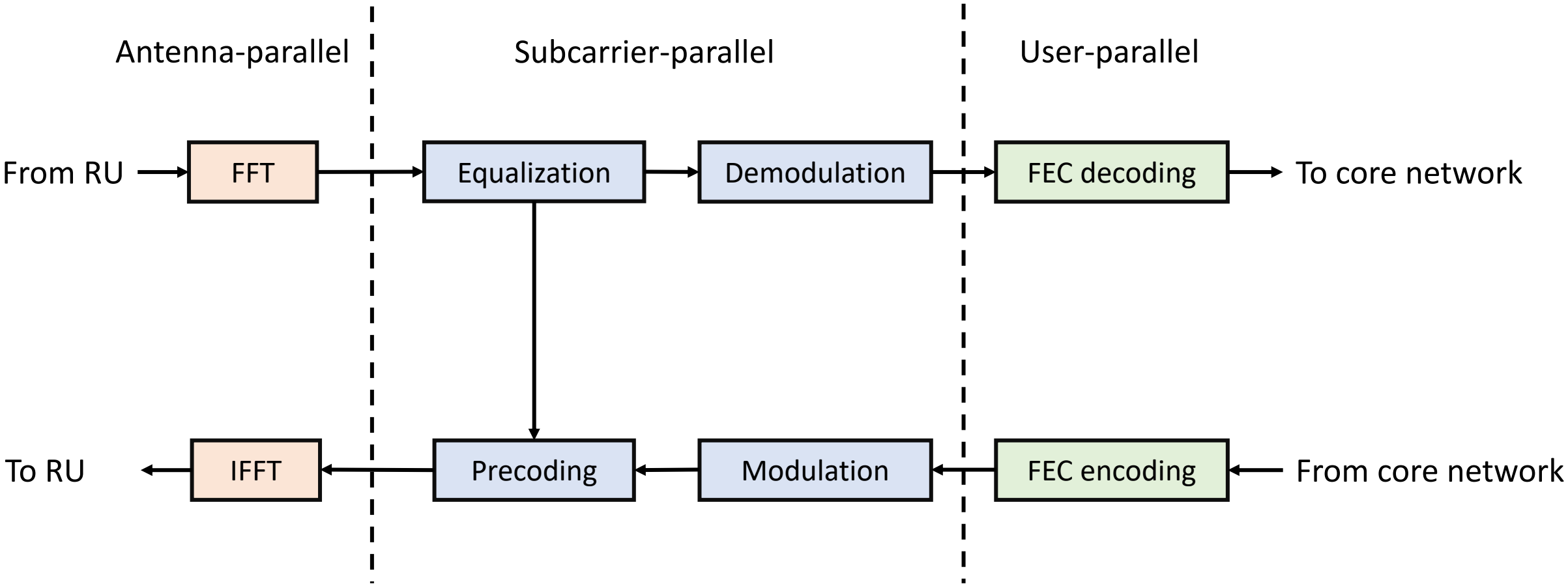
Reduce inter-server communication overhead

- Exploit RU features to deliver fronthaul data directly to servers instead of shuffling the data among servers in prior designs
- Delay shuffling until later in the pipeline when the data size is reduced

Reduce intra-server communication overhead

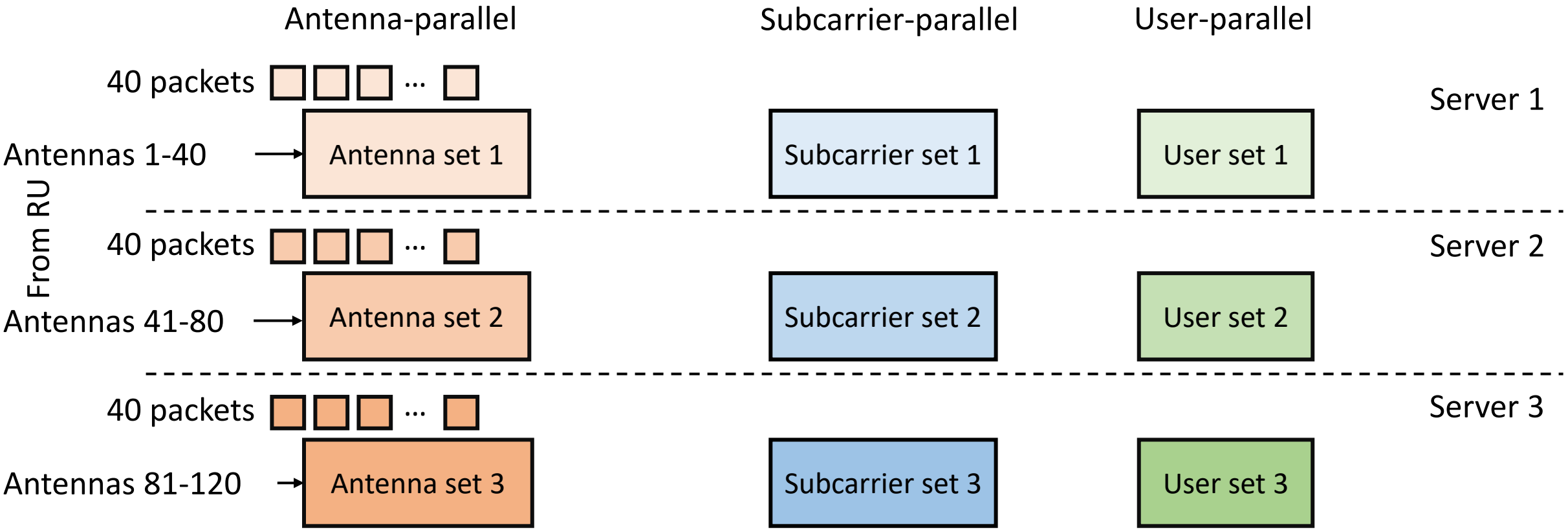
- Subcarrier-to-core affinity to minimize inter-core data movement
- Eliminate centralized task scheduling

Background: massive MIMO processing pipeline



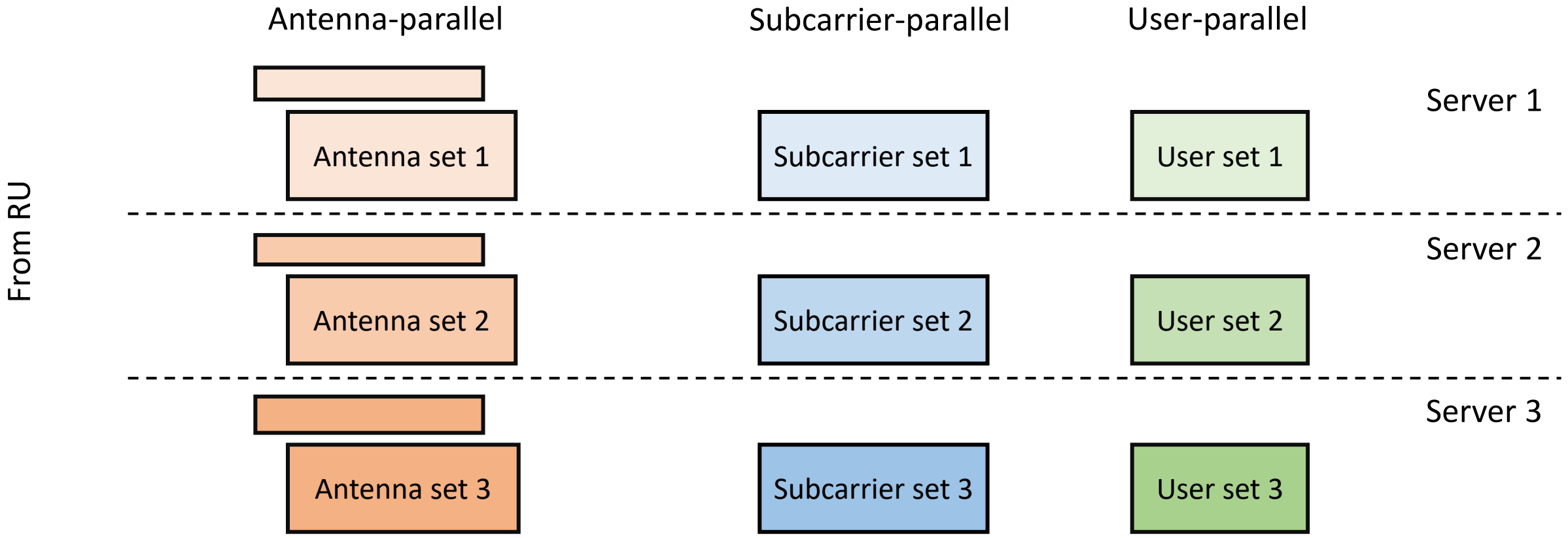
Data dependency between stages introduces communication overhead

Example: 120x30 MIMO, 1200 subcarriers



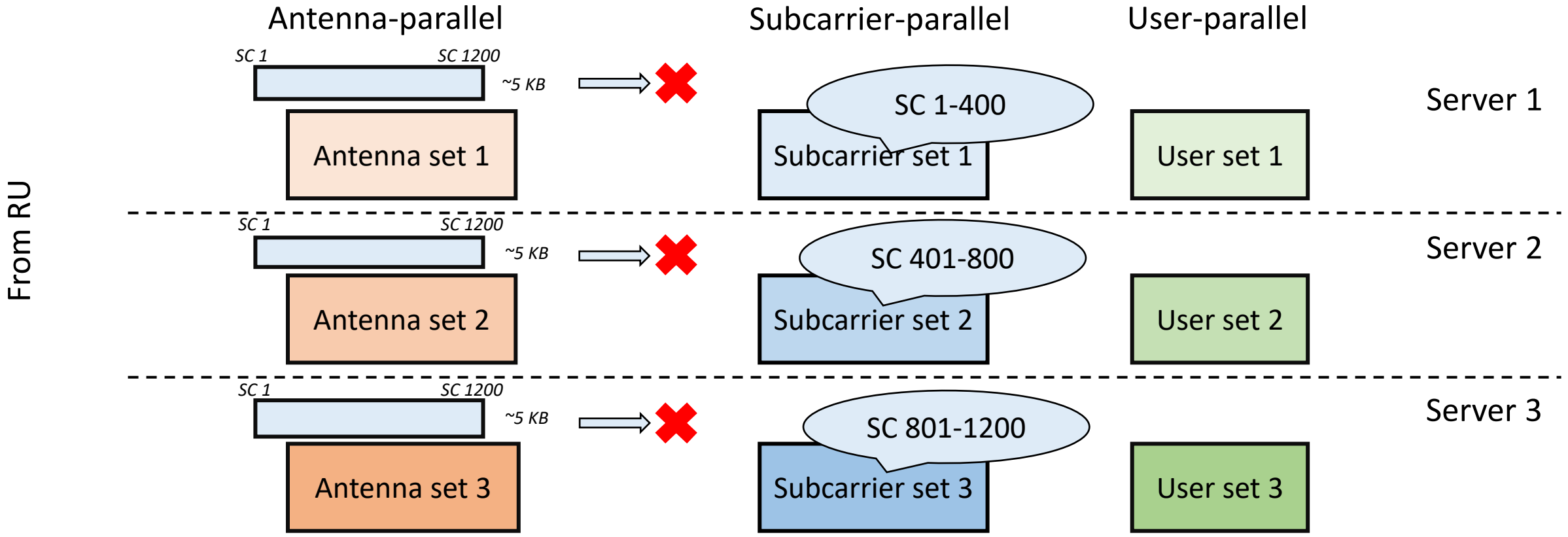
Data dependency between stages introduces communication overhead

Example: 120x30 MIMO, 1200 subcarriers



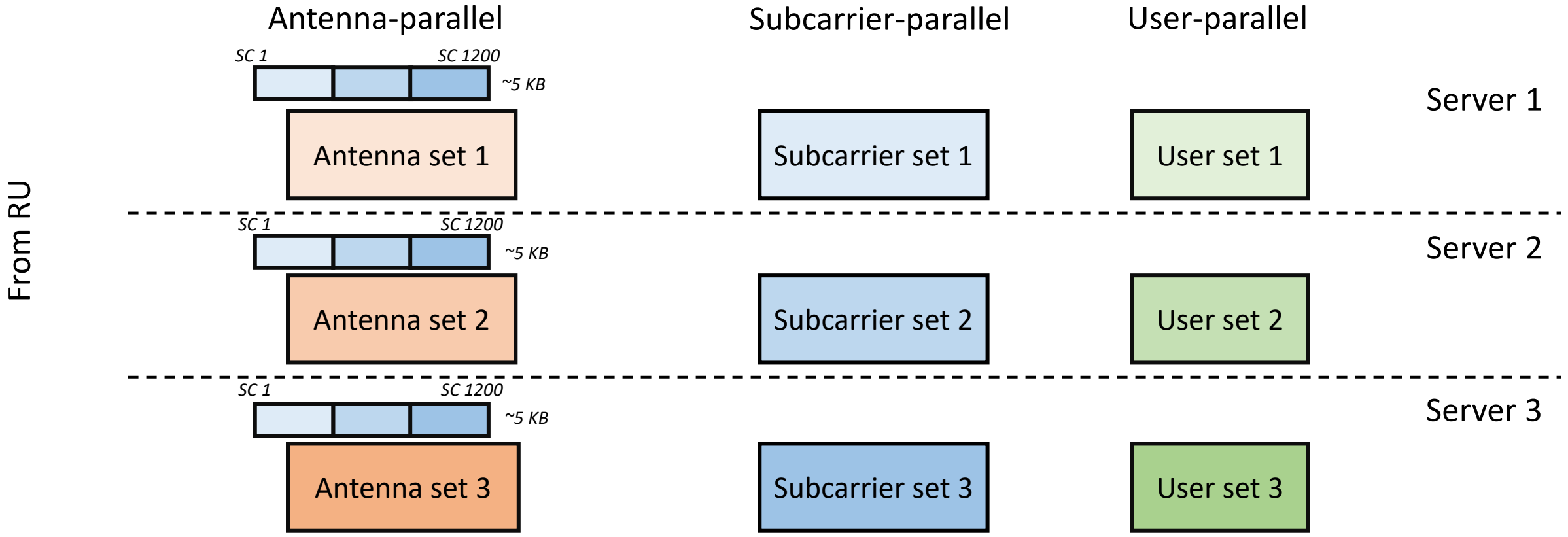
Data dependency between stages introduces communication overhead

Example: 120x30 MIMO, 1200 subcarriers



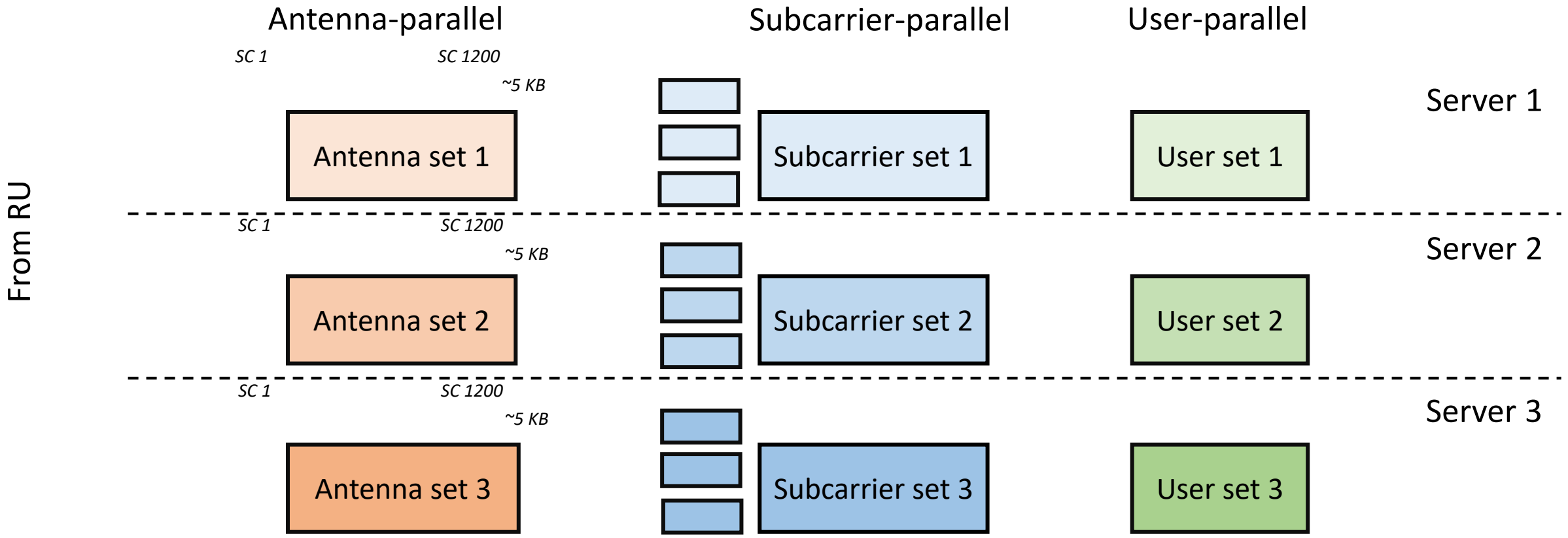
Data dependency between stages introduces communication overhead

Example: 120x30 MIMO, 1200 subcarriers



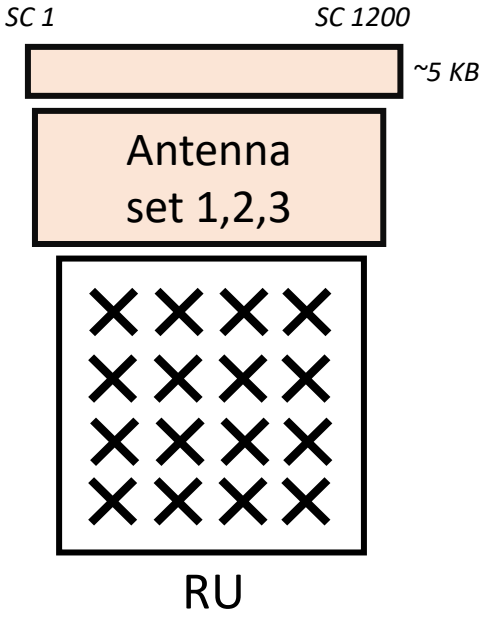
Data dependency between stages introduces communication overhead

Example: 120x30 MIMO, 1200 subcarriers



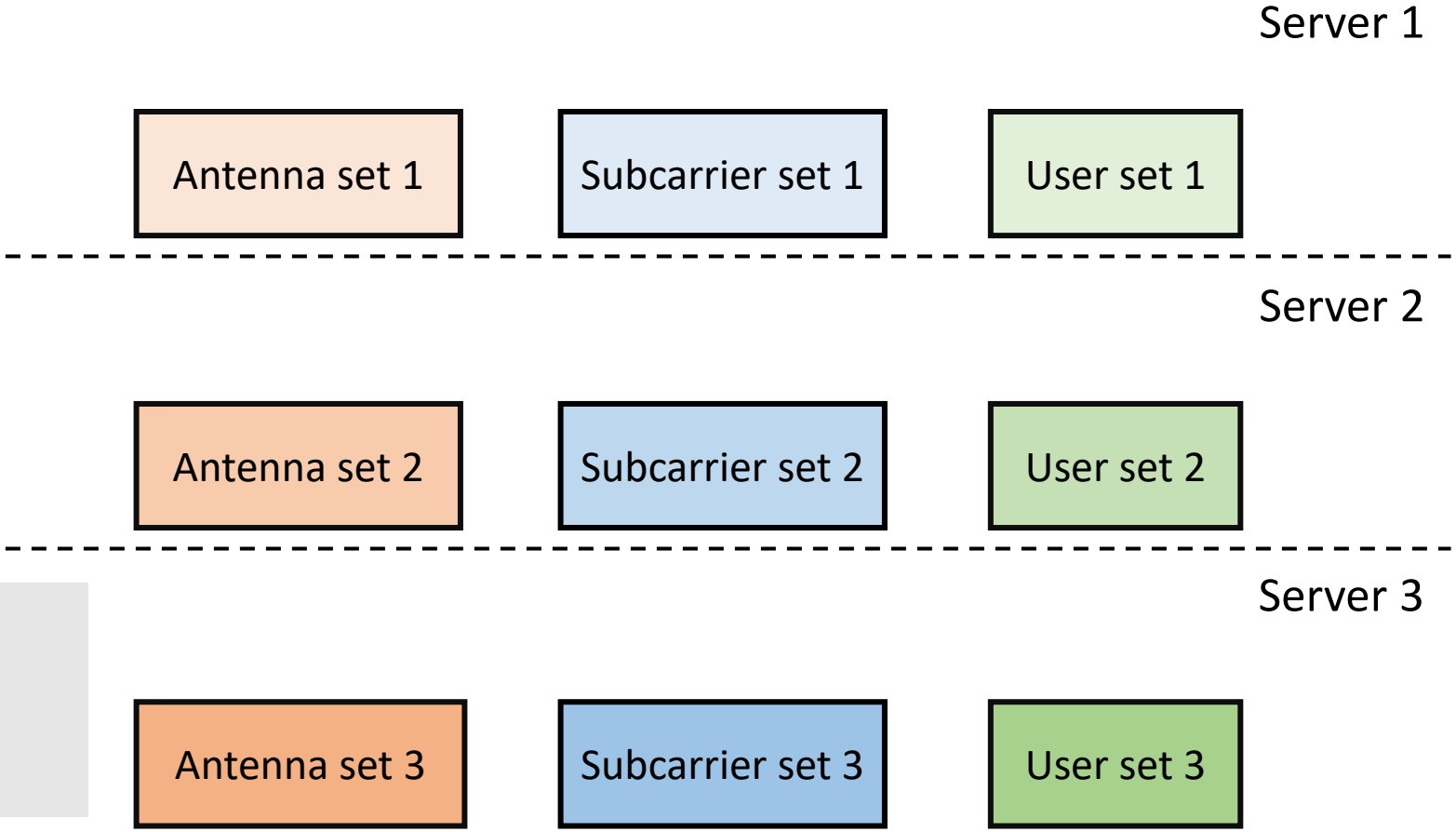
Scalability bottleneck: High rate (> 120 Gbps) of inter-server shuffling

Idea #1: Exploit modern RU features to avoid data shuffling

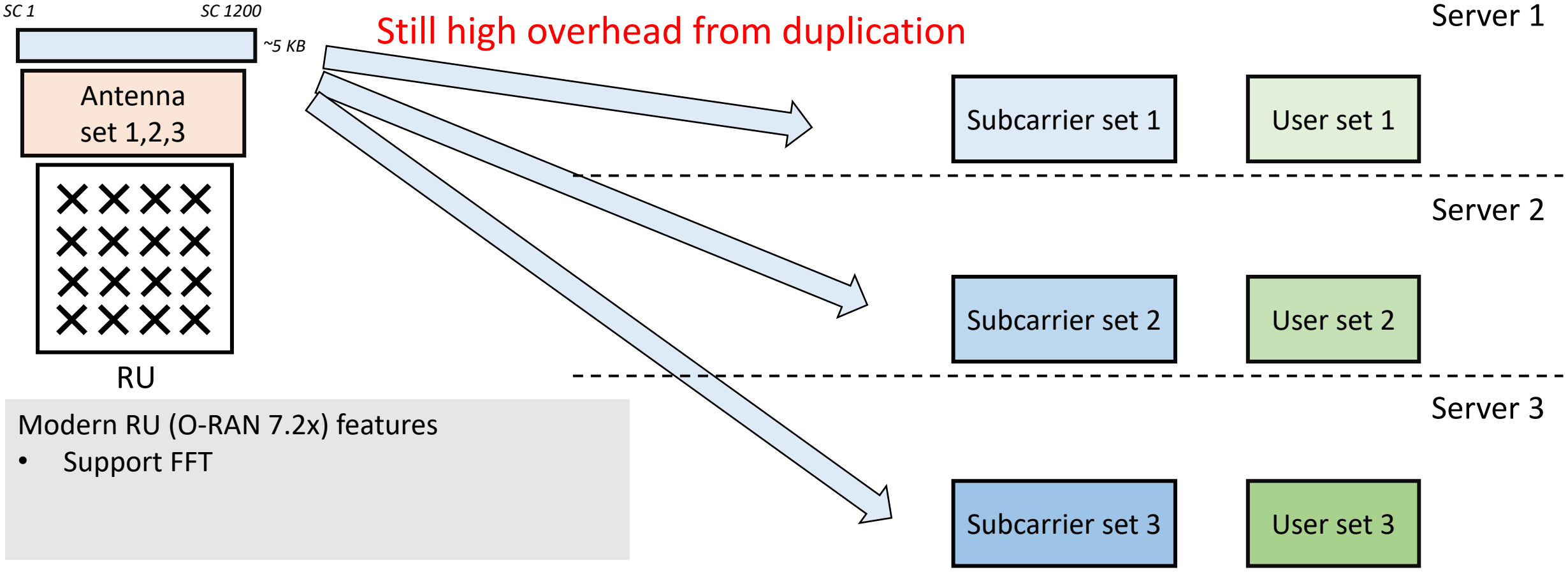


Modern RU (O-RAN 7.2x) features

- Support FFT



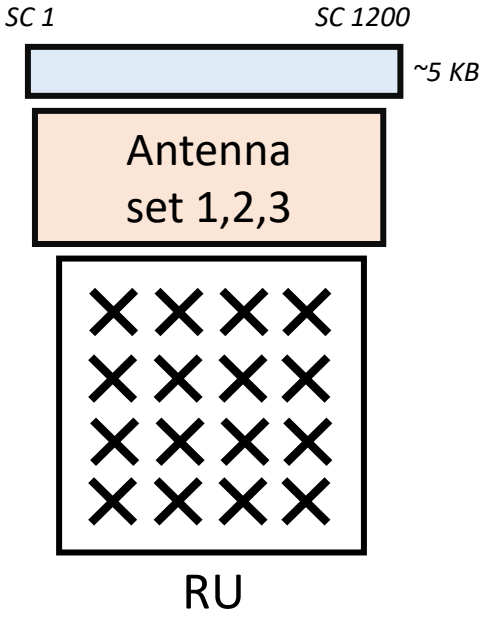
Idea #1: Exploit modern RU features to avoid data shuffling



Modern RU (O-RAN 7.2x) features

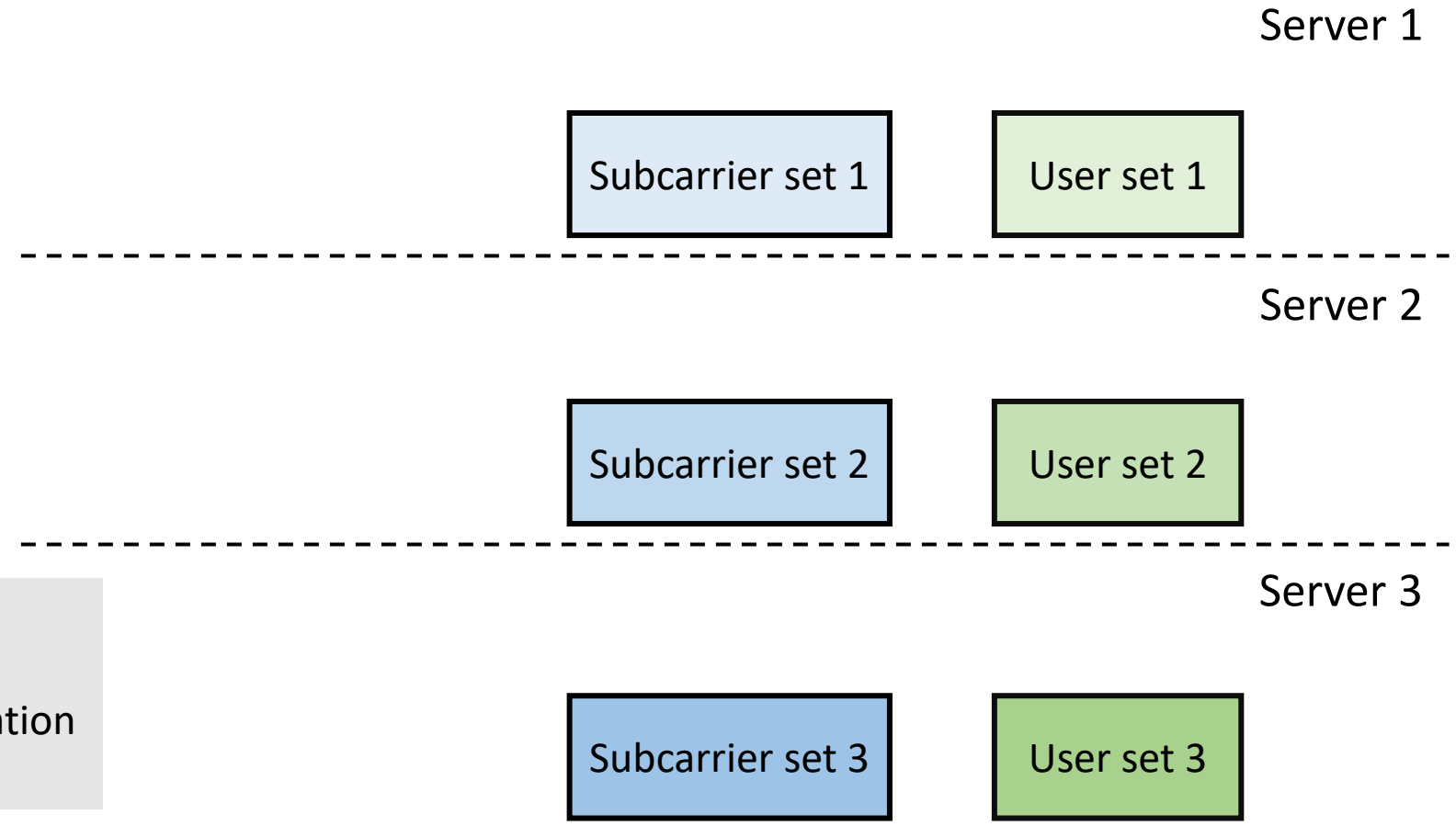
- Support FFT

Idea #1: Exploit modern RU features to avoid data shuffling

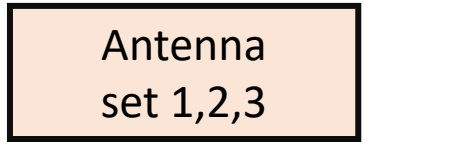


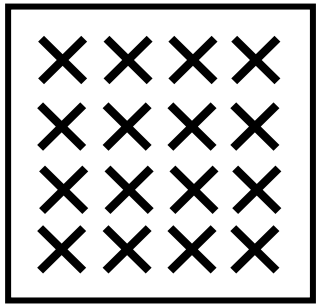
Modern RU (O-RAN 7.2x) features

- Support FFT
- Configurable fronthaul packet segmentation
  - Originally designed for MTU tuning



Idea #1: Exploit modern RU features to avoid data shuffling

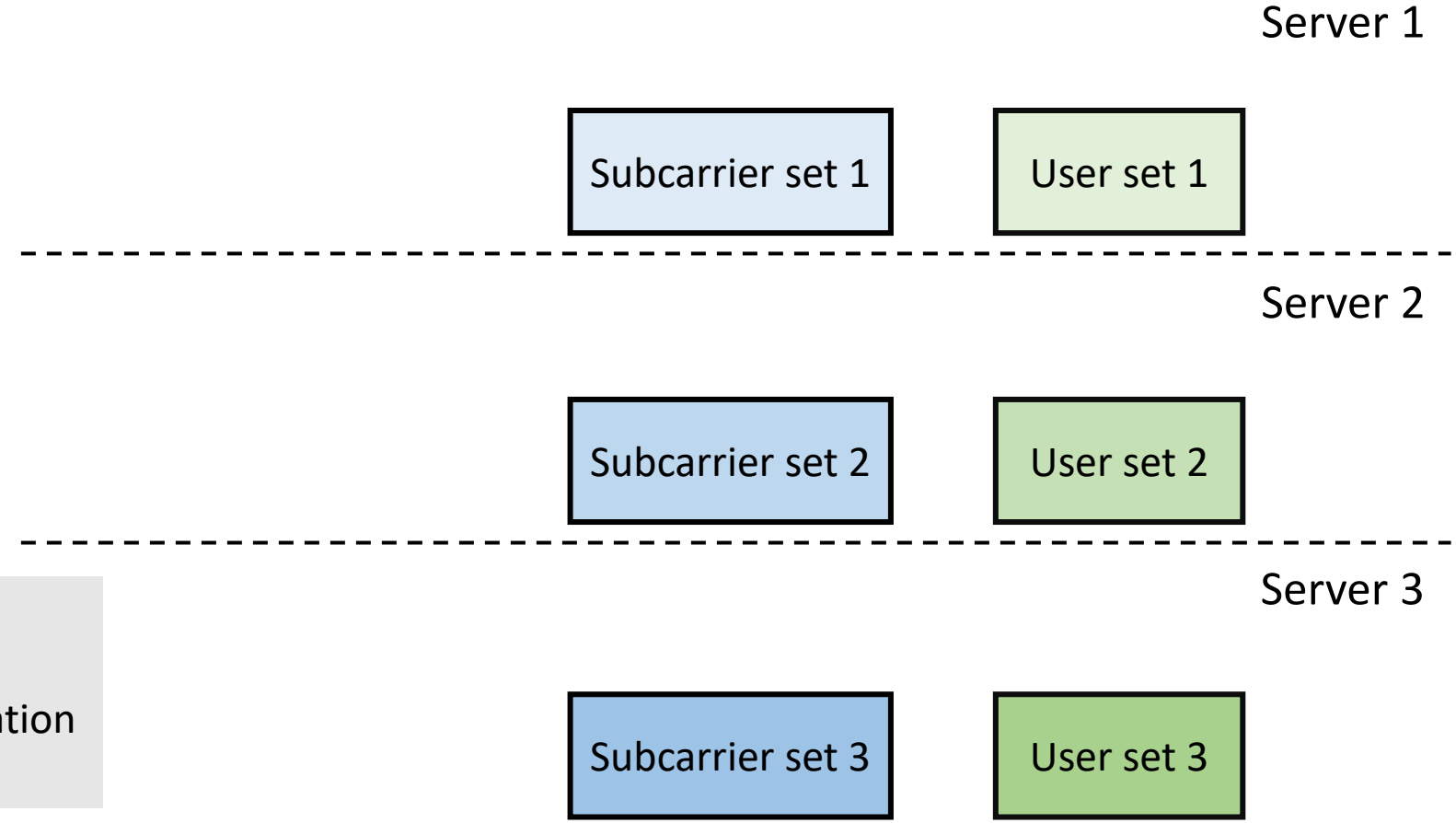




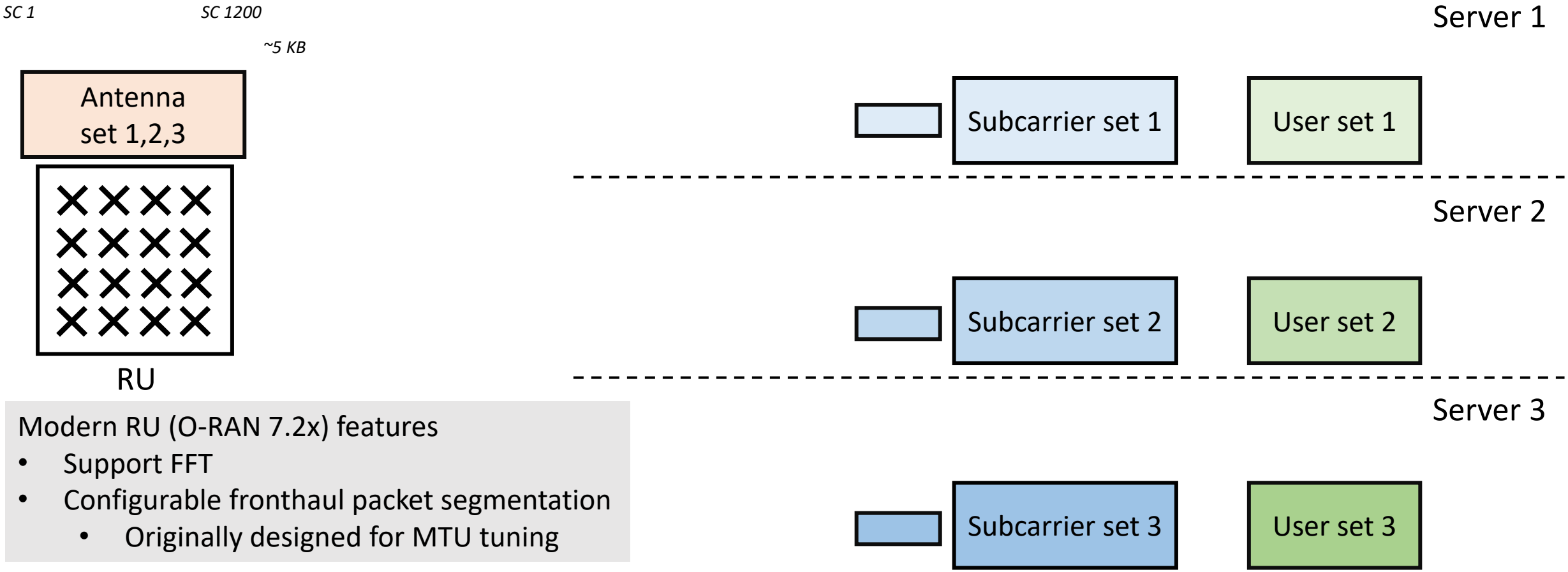
RU

Modern RU (O-RAN 7.2x) features

- Support FFT
- Configurable fronthaul packet segmentation
  - Originally designed for MTU tuning



Idea #1: Exploit modern RU features to avoid data shuffling

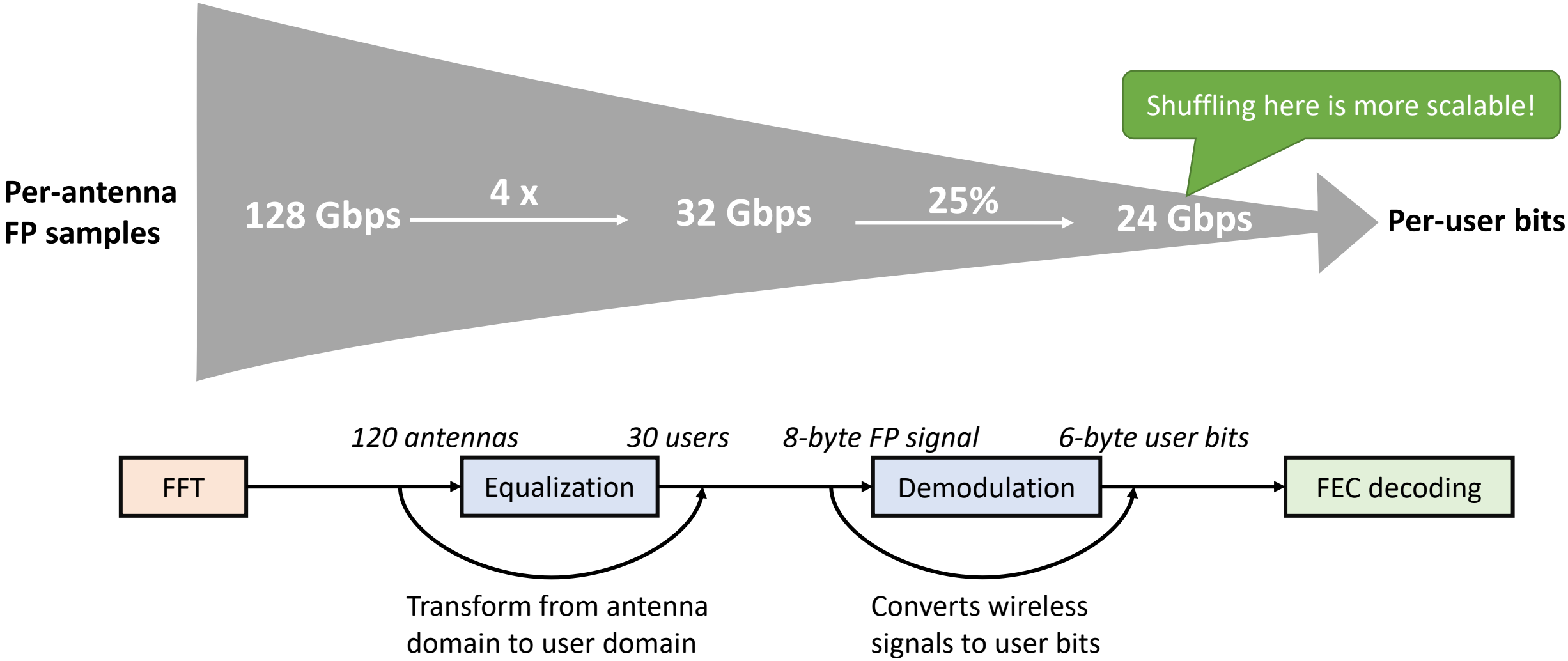


Modern RU (O-RAN 7.2x) features

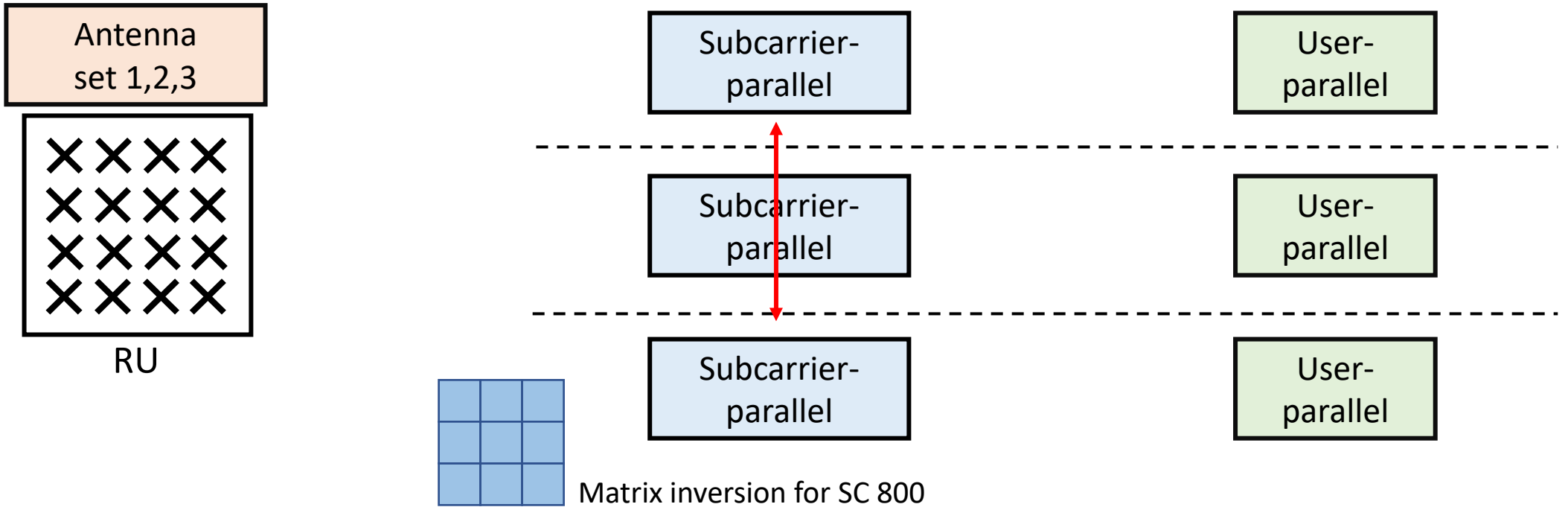
- Support FFT
- Configurable fronthaul packet segmentation
  - Originally designed for MTU tuning

Hydra eliminates fronthaul shuffling by leveraging modern RU features

Observation: the pipeline progressively reduces the data size

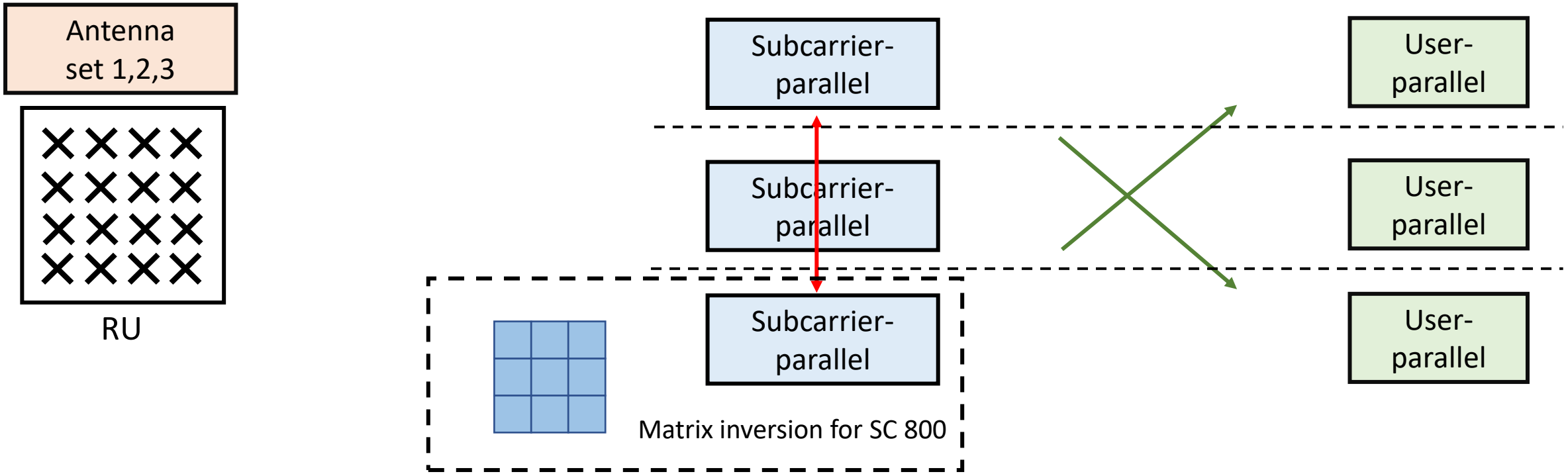


Intuitive parallelization can increase inter-server communication



- Maximizing parallelism makes sense when CPUs are weak (e.g., BigStation)
- Limits scalability due to high inter-server communication with large numbers of antennas and users

Idea #2: Affinitize subcarriers to a dedicated server



Shuffling only after the subcarrier-parallel stage: low overhead due to data size reduction

Evaluation setup

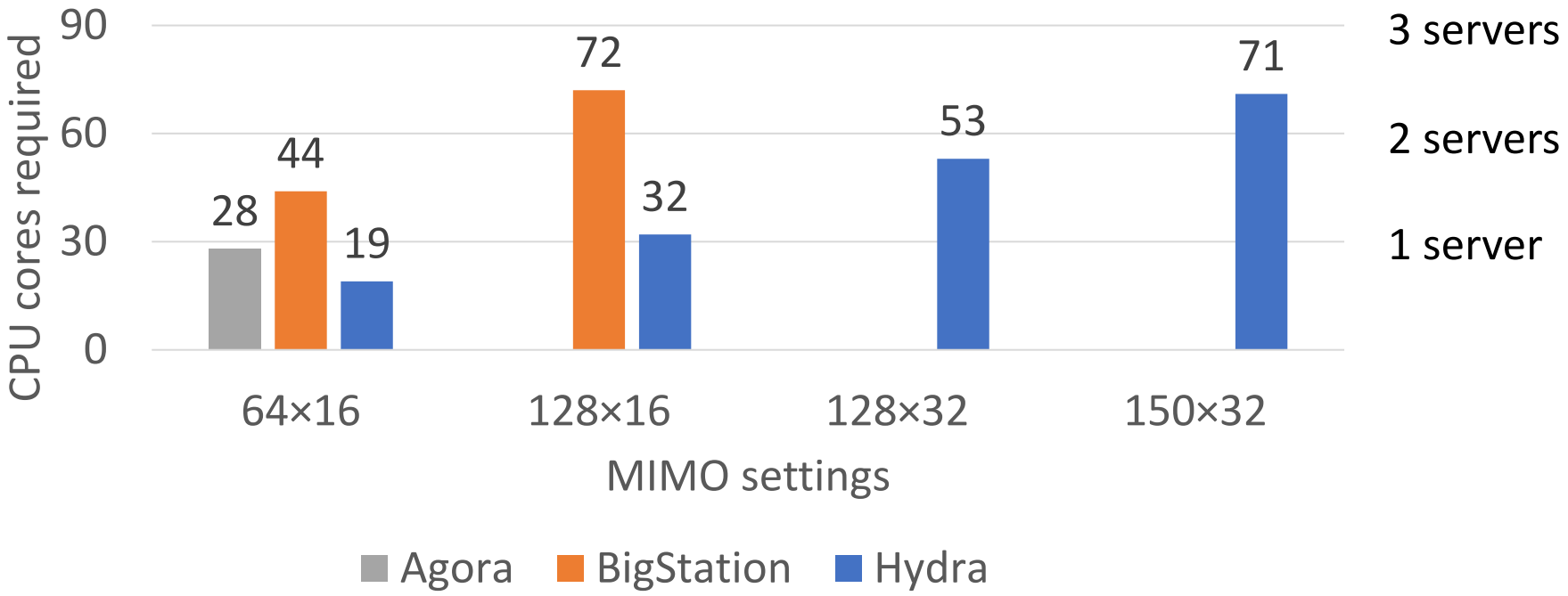
Hardware configurations

- Four commodity servers
- Each server has two 16-core CPUs, with AVX2 support
- 100 GbE NIC

Experiments were done with RU emulator

- Three servers for Hydra
- One server for RU emulator

Hydra is more scalable than existing solutions



Hydra supports more challenging MIMO settings

Experiment on more servers

- 27 servers in CloudLab (18 for Hydra, 9 for RU emulator)
- Hydra supports 256×32 MIMO (Uplink) with 18 servers

Conclusion: Hydra's massive MIMO processing is scalable

- We show that inter- and intra-server communication is a key scalability limiter in prior massive MIMO designs
- Hydra's scalability comes from
  - Using features of modern RUs in novel ways
  - Efficient computation partitioning
- Hydra supports 150×32 MIMO for the first time in software
- Hydra's scalability makes rapid development and deployment of 5G networks possible

Thank you!