



Software Architecture Improvements and FPGA Acceleration for OAI 5G NR gNB

Make OAI fast on a low-clock, older-core Xeon → then selectively replace critical PHY components with FPGA accelerators

Nung Yi Kuo

ISIP Lab · National Yang Ming Chiao Tung University (NYCU), Taiwan

OAI Software Architecture & E2E Optimization: Syed Ibad Ur Rahman

OAI Foundation Hands-on Workshop
North Carolina State University · Raleigh,
NC
13–14 October 2026

5G NR · O-RAN 7.2 · OAI · ACC100 · FPGA

What We Will Demonstrate at the Workshop

A compact story from software bottlenecks → architecture improvement → selective hardware acceleration

- 1 Why this test platform** Low-clock, older-core Xeon is a deliberate stress target: expose bottlenecks and target high-end-CPU-class performance without a CPU upgrade
- 2 Software improvement keys** Parallel DL processing, copy removal, ACC100 overlap, parallel FHI, and cross-slot UL scheduling
- 3 Measured impact on our platform** Same older Xeon: DL 371 → 847 Mbps average, 958 Mbps peak; UL 164 Mbps long-run
- 4 FPGA DL accelerator** Selectively replace the DL PDSCH compute chain; 1.4 Gbps E2E operation
- 5 FPGA UL accelerator** Selectively replace UL PUSCH CE/MMSE/demod/LLR; stable high-load operation

One message: use a constrained CPU to prove software efficiency first; then selectively offload critical PHY blocks to FPGA.

Why This Test Platform Matters — Low-Clock, Older-Core Xeon

A constrained E2E platform makes bottlenecks visible and shows what architecture optimization can recover before hardware offload

Why this CPU?

2 × Intel Xeon Gold 6138

2.0 GHz base clock
Low clock rate
Older core architecture
40 physical cores · HT off

Why use this platform?

Stress-test platform

100 MHz · 273 PRB · up to 4 DL layers

Tight slot deadlines expose CPU bottlenecks in precoding, modulation, memory movement, fronthaul processing, and UL scheduling.

Actual E2E setup with FPGA adapter



Same host used for OAI optimization and PCIe-connected FPGA offload

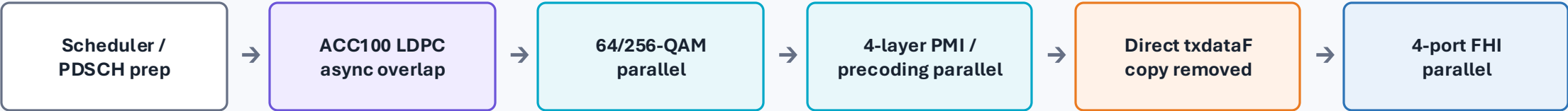
What can we get from this platform?

- Quantify before/after gains on the same constrained CPU
- Target high-end-CPU-class throughput without replacing the server
- Validate real E2E O-RAN operation and reuse the same host for ACC100 / FPGA offload

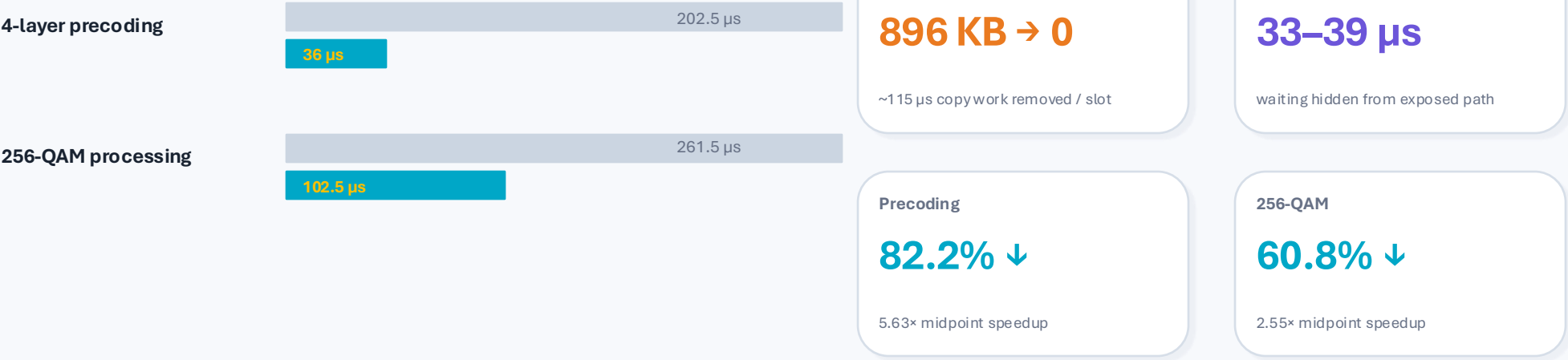
Motivation: use a low-clock, older-core Xeon to target high-end-CPU-class OAI performance through architecture efficiency before upgrading hardware.

Improvement Keys — Downlink Software Architecture

Reduce the critical-path work exposed on the same older Xeon platform



Measured critical-stage improvement



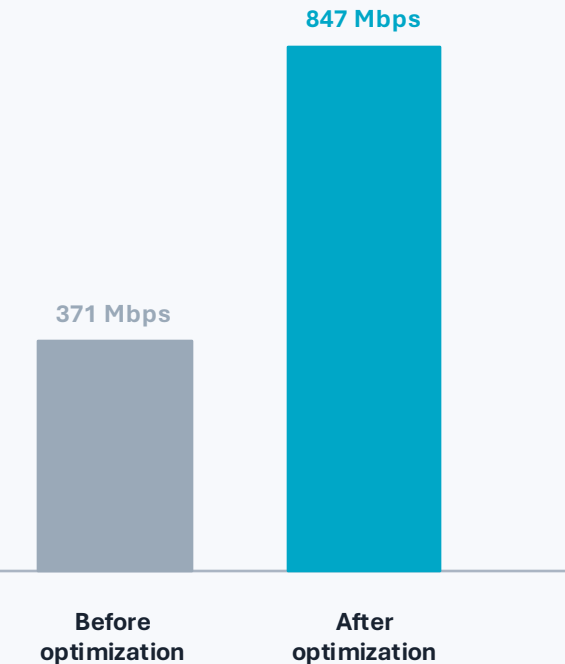
Key idea: keep OAI’s real-time structure, but remove serialization and memory overhead so the older Xeon has more processing headroom.

Measured Impact on Our Older Xeon Platform

Same 1 Gbps offered UDP load, same Xeon Gold 6138 platform: before vs after our architecture changes

Average receiver throughput

+476 Mbps · 2.28×



Peak 1-second DL

958 Mbps

optimized run

UDP loss

62% → 15%

75.8% relative reduction

1st HARQ retrans.

0.668% → 0.158%

76.4% relative reduction

Final DL errors

0

both runs ended with zero final DL errors

Delivered / offered

84.7%

847 / 1000 Mbps

4-layer operation

RI4

optimized high-load run

Real test evidence

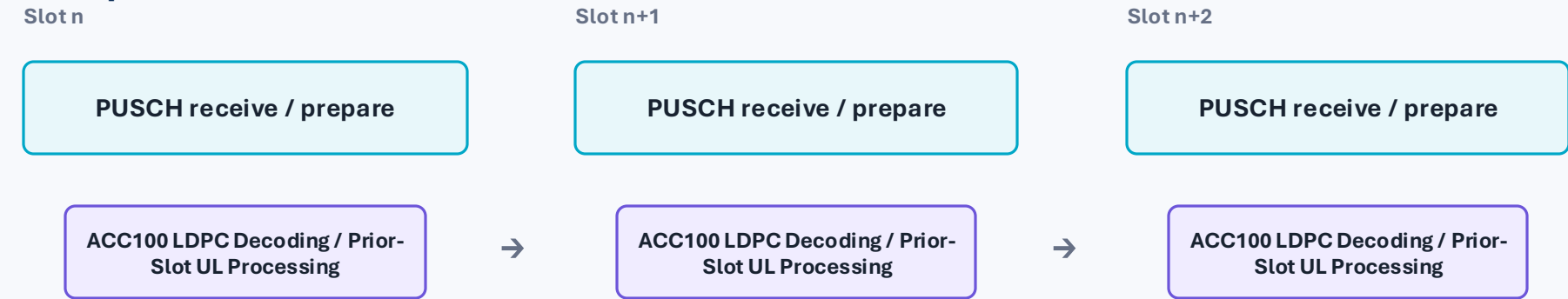


Same 2.0 GHz Xeon; no CPU upgrade. Internal before/after comparison only — not a claim against OAI’s published maximum. The 15% UDP loss is application-layer offered-load loss; final DL errors remained zero.

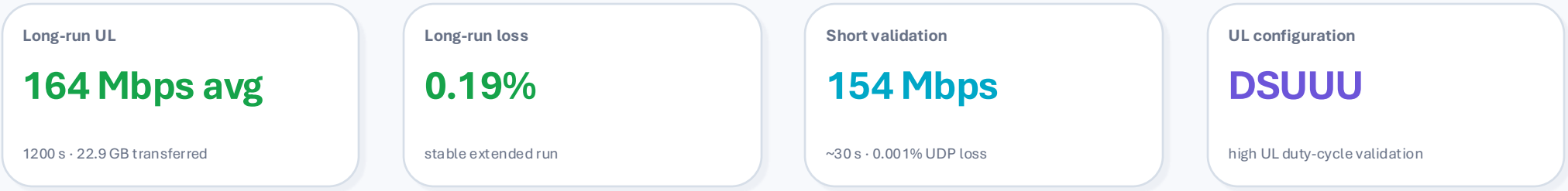
Improvement Key — Uplink Cross-Slot Pipeline

Use slot-level concurrency and ACC100 overlap to keep high-load UL processing inside real-time deadlines on the same Xeon

Cross-slot scheduling concept



Overlap next-slot PUSCH reception with prior-slot LDPC decoding

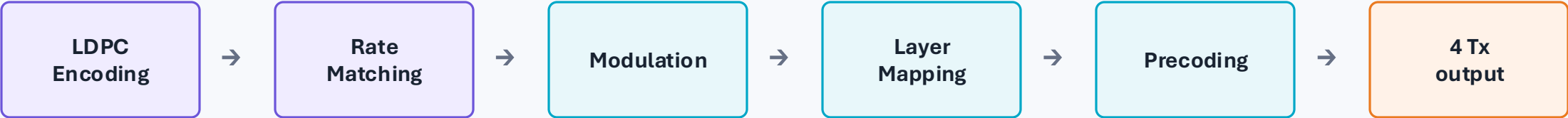


Result: stable high-load UL operation with accelerator-assisted cross-slot concurrency rather than a fully serial slot-by-slot path.

FPGA — Downlink PDSCH Accelerator

Next step: selectively replace high-complexity downlink PHY blocks with a deeply parallel FPGA chain

AMD / Xilinx VCU128 · PCIe Gen3
×16



CPU side keeps control-oriented tasks (e.g., CRC / DMRS handling); high-volume PDSCH compute is moved to FPGA.

E2E DL throughput

1.4 Gbps

4-layer · MCS25 · E2E

Worst-case FPGA slot

80–151 μs

RV0–RV3 · 273 RB · 4 layer · MCS27

Chip power

≈15.0 W

Vivado post-implementation estimate

Architecture

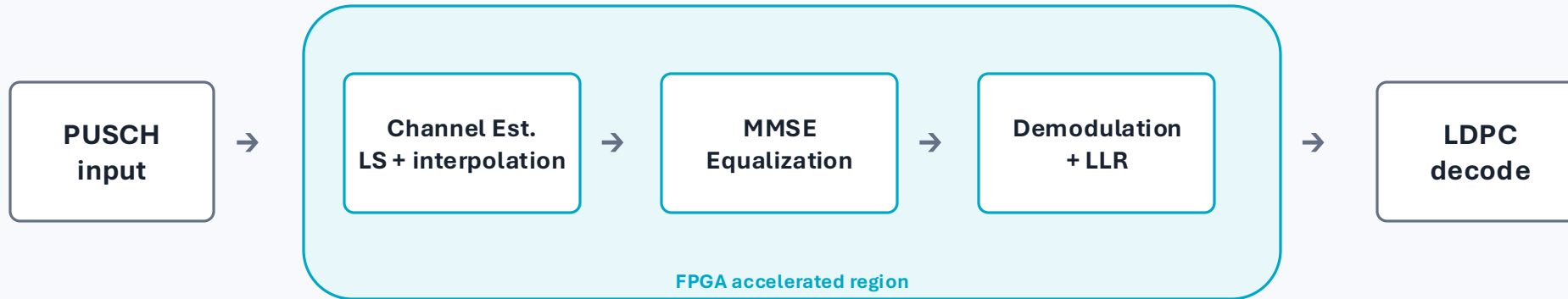
4-way CBP

parallel Code Block Processor

Hardware objective: deterministic high-throughput PDSCH processing while releasing CPU resources for protocol-stack work.

FPGA — Uplink PUSCH Accelerator

Next step: selectively move PUSCH channel-estimation / equalization / LLR compute out of the CPU receive path



E2E UL stability

≈200 Mbps

DSUUU · >30 min continuous

Integrated compute

≈28.5 μs

273 RB · 3 DMRS · 256QAM · 4Rx/1L

PUSCH path

400 → 300 μs

~100 μs end-to-end path saving

Pipeline

CE → MMSE

RB-level overlap hides CE latency

Measured decomposition after FPGA integration: ~220 μs PCIe transfer + ~30 μs FPGA compute + ~50 μs host packing/unpacking.

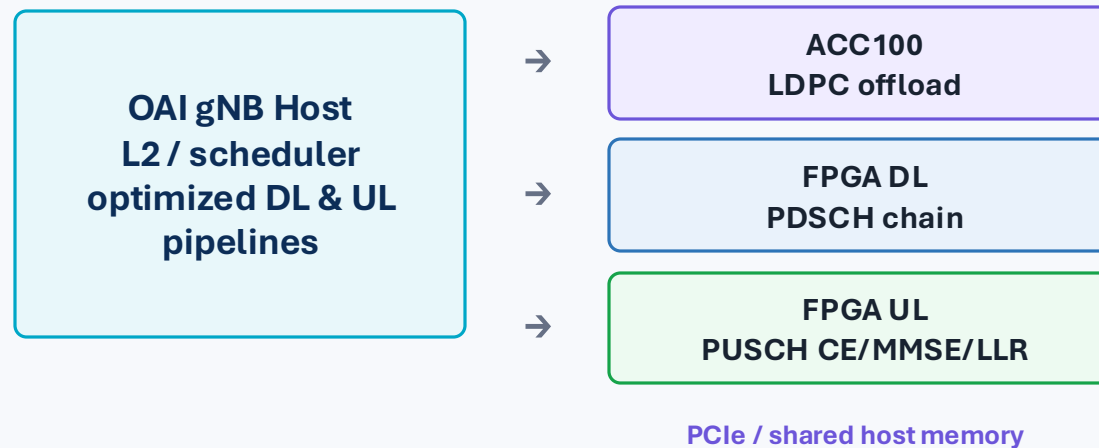
The bottleneck shifts from algorithm computation toward data movement — a clear target for the next optimization step.

Integrated Heterogeneous OAI gNB Architecture

Software decides and schedules; accelerators execute selected time-critical PHY work



Same OAI host selectively offloads critical blocks over PCIe



Design principle: make OAI efficient on existing Xeon hardware first; selectively replace only the critical PHY blocks while preserving OAI flexibility.



Key Takeaways

Our goal is high-end-class OAI throughput on a low-clock, older-core Xeon without replacing the server CPU. Software architecture improvements recover processing headroom; selective FPGA offload extends it further.

847 Mbps

SW DL avg · 2.0 GHz Xeon

958 Mbps

SW DL peak · 2.0 GHz Xeon

164 Mbps

SW UL avg · same Xeon · 20 min

1.4 Gbps

FPGA DL · MCS25

≈200 Mbps

FPGA UL · >30 min

Nung Yi Kuo

ISIP Lab · National Yang Ming Chiao Tung University (NYCU), Taiwan

OAI Software Architecture & E2E Optimization: Syed Ibad Ur Rahman

**OAI Foundation Hands-on Workshop
North Carolina State University · Raleigh, NC
13–14 October 2026**

Thank you · Questions & Discussion