

Energy-Oriented Acceleration of Memory-Polynomial DPD on a CGLA

Mapping a local-reuse complex-MAC reduction onto IMAX

Takuto Ando, Yasuhiko Nakashima / NAIST

1. Background

RF front-end compensation is strongly power constrained

2. Mapping

Expand the 15-term memory-polynomial DPD into a 33-stage stream

3. Evaluation

Compare FPGA, RTX 4090, and Jetson ARM-NEON on the same workload

4. Conclusion

Show an energy-oriented CGLA dataflow as a kernel-level study

Conclusion: RTX 4090 is the low-latency point, while the IMAX ASIC projection occupies the low-energy region.

Background: DPD is a complex-MAC workload with local history reuse

Problem Setting

Digital predistortion compensates transmitter nonlinear distortion

A memory-polynomial form repeatedly reuses a small coefficient set and sliding input history

A GPU provides low latency, but RF front-end use cases also require energy and PDP views

This talk treats the work as a kernel-level accelerator study, not a full transmitter integration

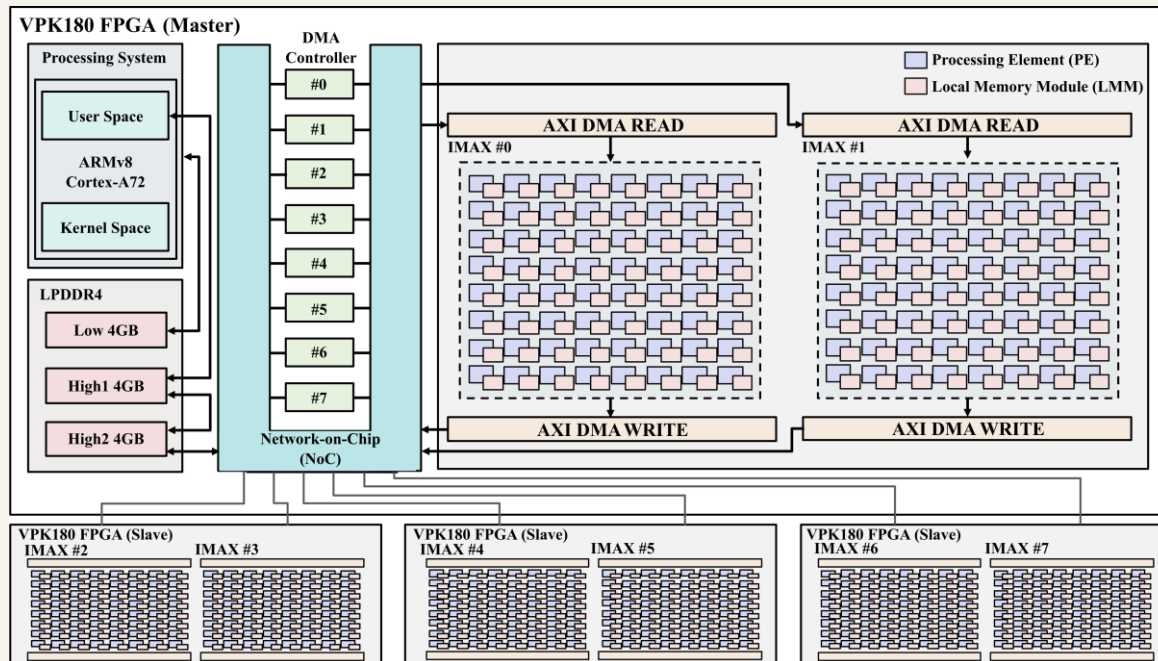
Research Questions

Can a 15-term complex-MAC reduction be placed on the one-dimensional IMAX PE/LMM pipeline?

Can the 120 B coefficient set and 5-tap history reuse be kept in local memory?

Does it make sense as an energy-oriented point rather than a latency winner?

IMAX/CGLA: stream execution with one-dimensional PEs and local memories



Properties Used for DPD

CPU-Grounded Linear Array: PEs and LMMs are interleaved in one dimension

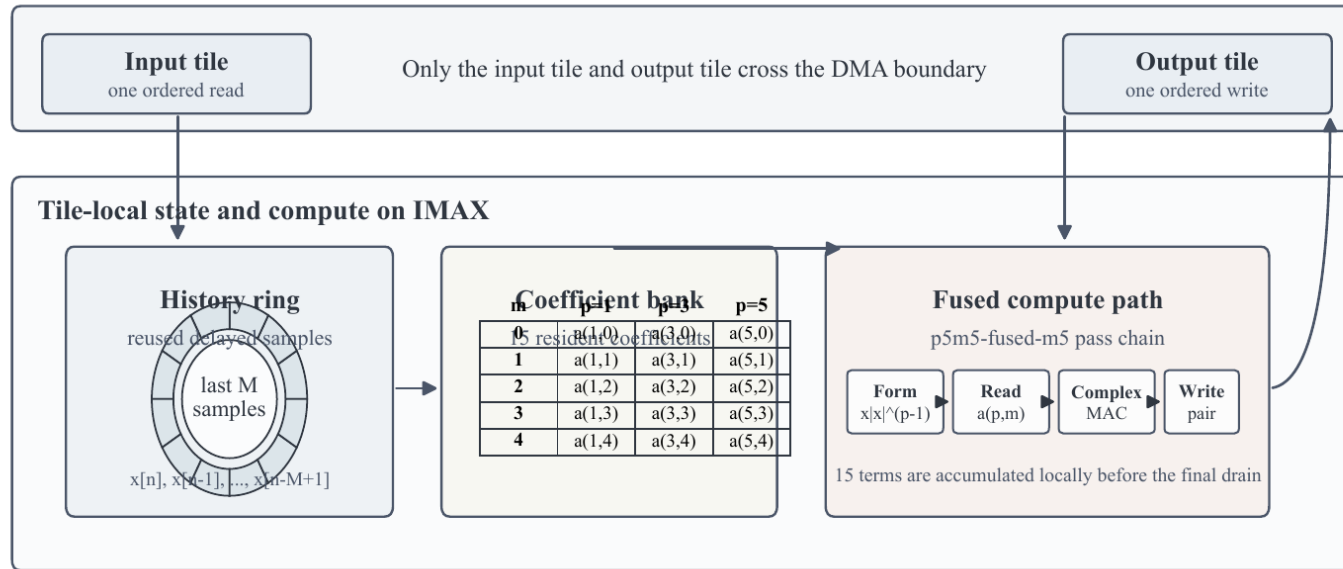
Coefficients, history, and partial sums are passed as streams

Explicit DMA and local reuse reduce external movement per operation

CGLA is treated as a programmable ASIC class

Proposed Mapping: turning 15-term DPD into a streaming complex-MAC

DMA-visible input/output stream



delay-state reuse across adjacent outputs the same 15 coefficients are reused for the whole tile partial sums remain local until the final output write

Reading Points

(P,M)=(5,5) odd-order memory-polynomial

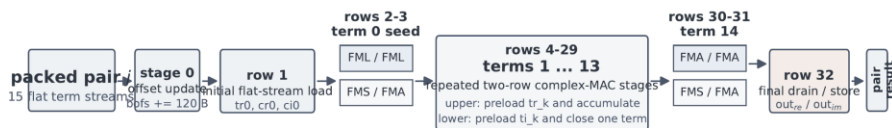
Keep the 120 B coefficient set in local memory

Advance the 5-tap history over 32-sample tiles

DPD PE Pipeline

Map one packed pair to a 33-PE streaming schedule.

Generated 33-PE DPD pipeline for one packed pair
offset update, initial load, 15 overlapped complex terms, and final store



How to read this figure

Entry

PE0 updates the 120 B pair stride; PE1 performs the initial flat-stream load.

Compute chain

PE2-PE31 advance 15 complex terms with a two-PE pattern per term.

Drain

PE32 receives the final partial result and stores out_re / out_im.

Scope

33 PEs are used in the 64-PE array; final placement can still depend on compilation.

Reading Points

PE0 offset update, PE1 load

PE2-PE31 implement the 15-term complex-MAC reduction

PE32 performs the final store

Evaluation Setup: compare the same single-precision workload

対象	内容	位置づけ
IMAX FPGA	VPK180 prototype / 145 MHz	実測
IMAX ASIC projection	28 nm frequency/power model	投影
RTX 4090	custom CUDA	低遅延 baseline
Jetson AGX Orin	ARM-NEON	edge CPU baseline

Evaluation Scope

Separate kernel latency from modeled energy

Evaluate the mappability of the DPD reduction kernel, not full transmitter integration

Main Result: GPU for latency, IMAX projection for energy

IMAX FPGA E2E

72.101 ms

prototype measured

IMAX FPGA kernel

8.565 ms

kernel-only measured

IMAX ASIC E2E

4.28 ms

28 nm projection

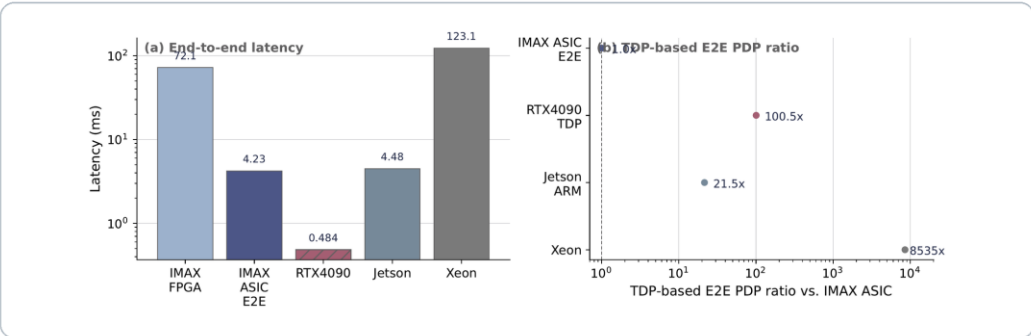
RTX 4090 E2E

0.484 ms

lowest latency

DPD Platform Comparison

Read latency and TDP-based E2E PDP as separate axes of evidence.



Left panel

End-to-end latency rows combine measured and projected entries.

Right panel

PDP is normalized against the IMAX ASIC E2E row using TDP-based power.

Axis policy

The latency axis and PDP-ratio axis use log scale; distances are multiplicative.

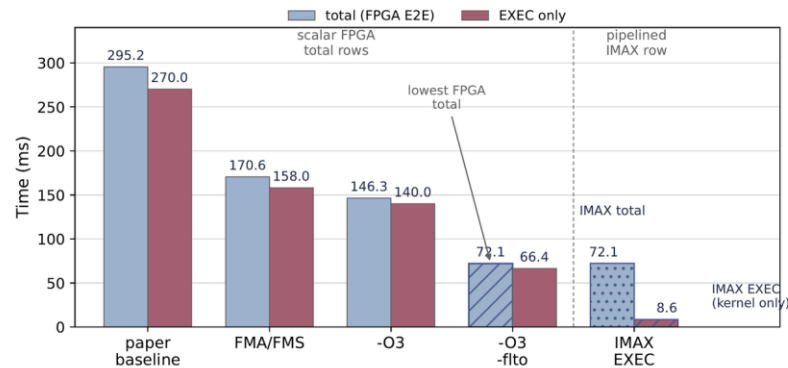
Guardrail

Kernel-only EXEC energy is reported separately in the following table.

Under the stated power assumption, the IMAX projection reduces modeled E2E energy per batch by 98.7x.

DPD Ablation

Separate prototype total time from the IMAX execute phase.



Read order

1 scalar rows -> 2 optimized scalar row -> 3 IMAX total and EXEC row.

Scope

The total row includes transfer, register setup, and ARM host preparation.

Evidence

IMAX total is 72.101 ms, while the isolated EXEC phase is 8.565 ms.

Guardrail

Use the total row for E2E timing and the EXEC row only for kernel-only timing.

Reading Points

Read not only the performance gap, but also which cuts preserve correctness

Use this to reason about kernel boundaries, DMA, and reuse

NMSE improvement

26.1 dB

controlled synthetic RF validation

ACLR improvement

26.0 dB

same 15-term form

How to Read This Result

These numbers validate the same 15-term form as a predistortion method, not as an accelerator comparison

The claim is that the DPD reduction kernel can be expressed as an energy-oriented CGLA dataflow

Limitations

- E2E latency is slower than RTX 4090
- Full transmitter integration has not been performed
- The ASIC projection depends on the stated frequency/power model

Next Steps

- Move to RF-chain integration that includes I/O and kernel boundaries
- Increase inter-tile reuse and DMA overlap
- Extend the method to communication-domain streaming reductions beyond DPD

Takeaways

Mapped the 15-term memory-polynomial DPD reduction to a 33-stage IMA stream

Positioned latency and energy using FPGA measurements and a 28 nm ASIC projection

GPU is the latency winner; IMA should be read as a sub-watt energy-oriented accelerator point

Full transmitter integration is kept explicit as future work