

2026 VLSI Review

서울대학교 전기정보공학부 박사과정 이재성

Session 27 Next-Generation Wireless Baseband Processors

이번 2026 IEEE/JSAP Symposium on VLSI Technology & Circuits의 Circuits Session 27은 Next-Generation Wireless Baseband Processors를 주제로 구성되었다. 본 글에서는 #27-1 tiled beamspace training/detection processor와 #27-2 multi-subcarrier massive MU-MIMO-OFDM detector를 중심으로 정리한다. 두 연구는 모두 mmWave massive MIMO에서 발생하는 높은 연산량과 메모리-데이터 이동 비용을 줄이기 위해 beamspace의 공간적 sparsity를 이용한다는 공통점이 있다.

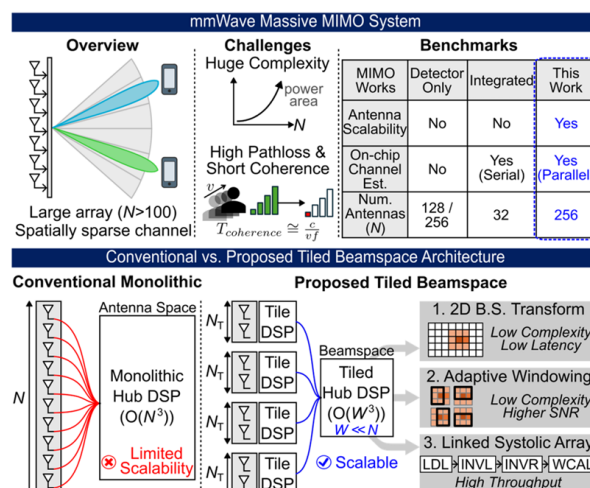
#27-1 본 논문은 mmWave massive MIMO의 training과 detection을 하나의 칩에서 처리하는 tiled beamspace processor를 제안한다. Massive MIMO는 base station에 수백 개 antenna를 사용하므로 array 규모가 커질수록 pilot sequence가 길어지고, channel estimation 이후의 matrix processing 복잡도와 모든 antenna chain을 단일 DSP로 연결하는 routing 부담이 급격히 증가한다. 특히 antenna-space 방식에서는 signal energy가 넓은 antenna dimension에 분산되어 있으므로 channel estimation의 covariance 계산과 weight calculation이 큰 차원의 matrix를 대상으로 수행된다. 저자들은 mmWave channel이 방향성으로 인해 spatially sparse하다는 특성에 주목하고, spatial FFT를 통해 antenna-space signal을 beamspace로 변환하면 에너지가 소수 beam bin에 집중된다는 점을 이용한다.

제안 구조의 핵심은 대규모 256-antenna array를 16개의 tile로 분할하고 각 tile이 16개 antenna를 담당하도록 한 뒤, local tile DSP와 hub DSP에 걸친 2-stage 2D beamspace transform을 수행하는 것이다. 첫 번째 intra-tile FFT 이후 각 tile에서는 폭 $W=4$ 의 window만으로도 90% 이상의 signal energy를 포착할 수 있고, 두 번째 inter-tile FFT를 거치면 전체 2D beam map에서 에너지가 다시 집중되어 $W_1 \times W_2 = 3 \times 3$, 즉 $D=9$ 의 reduced dimension으로 축소된다. 이후 adaptive trimming을 통해 이를 2×3 또는 3×2 로 추가 줄일 수 있다. 이에 따라 원래 $N=256$ 차원에서 수행되던 training과 detection을 훨씬 작은 D 차원으로 축소할 수 있으며, channel estimation의 p 와 R 계산은 각각 reduced dimension에서 진행되고 weight $w=R^{-1}p$ 계산의 $O(D^3)$ 부담도 크게 낮아진다.

이 접근의 의미는 단순히 FFT를 앞단에 추가하는 것이 아니라, 이후의 matrix inversion과

vector-matrix multiplication이 다루는 problem size 자체를 축소한다는 데 있다. 논문은 full antenna-space baseline과 비교했을 때 2D beamspace transform이 training area를 약 3 orders of magnitude, detection area를 약 1.5 orders of magnitude 줄일 수 있음을 보인다. 즉, beamspace는 통신 성능을 유지하면서도 후단 digital baseband가 실제로 처리해야 하는 state dimension을 줄여주는 algorithm-hardware co-design의 핵심 수단으로 사용된다.

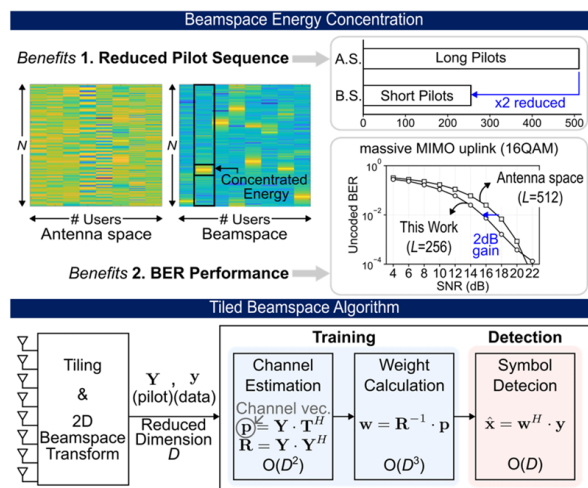
Reduced dimension을 얻은 뒤에는 multi-user channel estimation을 효율화하기 위한 추가 최적화가 적용된다. Covariance matrix R 은 Hermitian symmetric이므로 저자들은 lower triangular element만 계산하여 estimator area를 44% 줄였고, ± 1 Walsh-Hadamard pilot sequence를 이용하는 p estimator에서는 일반 multiplier를 mux-add 구조로 대체하여 추가적으로 54%의 area를 절감하였다. 또한 3×3 beam window를 항상 유지하지 않고 p의 energy distribution을 이용해 2×3 또는 3×2 로 trimming하는 adaptive windowing을 적용한다. 이 방식은 fixed 3×3 window 대비 area를 34% 줄이면서, 단순 fixed 2×3 trimming과 비교하면 dominant beam 방향을 따라 window를 선택할 수 있어 약 1.8 dB의 SNR 이득을 제공한다.



[그림 1] mmWave massive MIMO 처리 문제와 proposed tiled beamspace architecture

그림 1은 저자들이 해결하려는 scalability 문제와 이를 위한 tiled beamspace 구조를 보여준다. 기존 antenna-space 기반 monolithic architecture에서는 수백 개 antenna의 데이터를 하나의 hub DSP로 모아 처리해야 하므로 routing과 matrix-processing complexity가 array 크기에 따라 빠르게 증가한다. 제안 구조는 256개 antenna를 16개의 tile로 분할하고, 각 tile의 local DSP에서 먼저 spatial FFT를 수행한 뒤 hub DSP에서 inter-tile processing을 이어가는 방식으로 timing-critical data movement를 지역화한다. 이처럼 physical array tiling과 beamspace 변환을 함께 설계함으로써 단일 대형 DSP에 집중되는 데이터 경로를 분산시키고, 후단 연산이 처리해야 하는 유효 dimension도 함께 줄인다. 특히 channel estimation뿐 아니라 weight calculation과 symbol detection까지 on-chip에

통합하여 training과 detection을 하나의 scalable dataflow로 연결한 점이 기존 detector-only 구조와 구별된다.



[그림 2] beamspace energy concentration과 reduced-dimension training/detection algorithm

그림 2에서는 beamspace 변환이 실제 algorithm complexity 감소로 이어지는 과정을 확인할 수 있다. Antenna-space에서는 signal energy가 넓은 antenna dimension에 분산되지만 spatial FFT 이후에는 소수 beam bin에 집중되므로, 더 짧은 pilot sequence로도 유효 channel 정보를 확보할 수 있고 BER 측면에서도 유리하다. 해당 논문에서는 intra-tile FFT 후 각 tile에서 $W=4$ window를 선택하고, hub DSP의 두 번째 inter-tile FFT를 통해 최종 2D beam map에서 $W1 \times W2 = 3 \times 3$, 즉 $D=9$ 의 작은 영역만 남긴다. 이에 따라 channel estimation의 $\mathbf{p}$ 와 $\mathbf{R}$ 계산, $O(D^3)$ 의 weight calculation, 그리고 payload symbol detection을 모두 reduced dimension에서 수행할 수 있다. 논문이 제시한 결과에서는 full antenna-space processing 대비 training area가 약 3 orders of magnitude, detection area가 약 1.5 orders of magnitude 감소하며, beamspace의 energy concentration 자체가 후단 hardware 규모를 줄이는 핵심 역할을 한다.

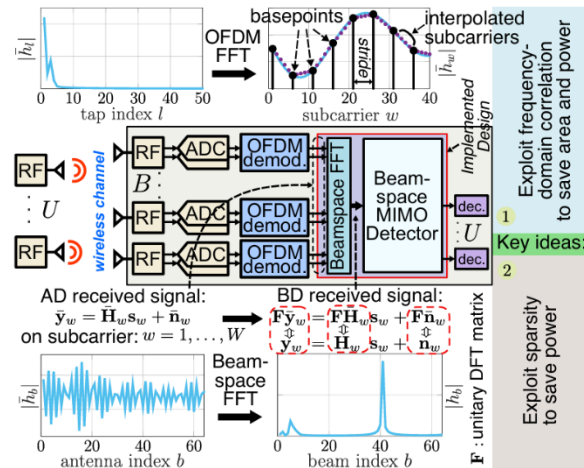
이 reduced-dimensional training flow에서는 $\mathbf{p}$ estimator와 covariance $\mathbf{R}$ calculator가 pilot sequence를 병렬 처리한다. $\mathbf{R}$ 은 Hermitian symmetric이므로 lower triangular element만 계산하여 estimator area를 44% 줄이고, ± 1 Walsh-Hadamard pilot을 이용하는 $\mathbf{p}$ estimator에서는 multiplier를 mux-add 구조로 대체하여 추가적으로 54%의 area를 절감한다. 또한 그림 2의 3×3 window를 항상 고정하지 않고 $\mathbf{p}$ 의 energy distribution을 이용해 2×3 또는 3×2 로 trimming하는 adaptive windowing을 적용한다. 이 방식은 fixed 3×3 window 대비 area를 34% 줄이면서 fixed 2×3 trimming보다 약 1.8 dB 높은 SNR을 제공한다. 이후 가장 큰 hardware bottleneck인 weight calculation은 $\mathbf{R} = \mathbf{L}\mathbf{D}\mathbf{L}^H$ decomposition, triangular matrix inversion $\mathbf{L}^{-1}$, inverse matrix assembly $\mathbf{R}^{-1} = (\mathbf{L}^{-1})^H \mathbf{D}^{-1} \mathbf{L}^{-1}$, WCAL의 네 단계로 구성된다. 저자들은 stage 사이에 전체 matrix를 저장하는 alignment buffer

를 없애고 LDL-INVL-INVR-WCAL을 직접 연결하는 linked systolic dataflow를 사용하여 intermediate storage와 stall을 동시에 줄였다.

이 linked systolic array는 단순한 PE count 감소보다 data alignment와 intermediate storage를 없애는 데 초점이 있다. LDL-INVL direct linking만으로 baseline 대비 area 23%, latency 22%를 줄였으며, INVR와 WCAL까지 포함한 end-to-end linked organization에서는 baseline 대비 area 66%, power 76% 감소와 20% latency 개선을 달성하였다. 즉, beamspace가 algorithm dimension을 줄이고 linked systolic dataflow가 그 작은 matrix를 불필요한 buffering 없이 처리하도록 구성되어, algorithmic sparsity와 architecture-level streaming이 서로 보완되는 형태이다.

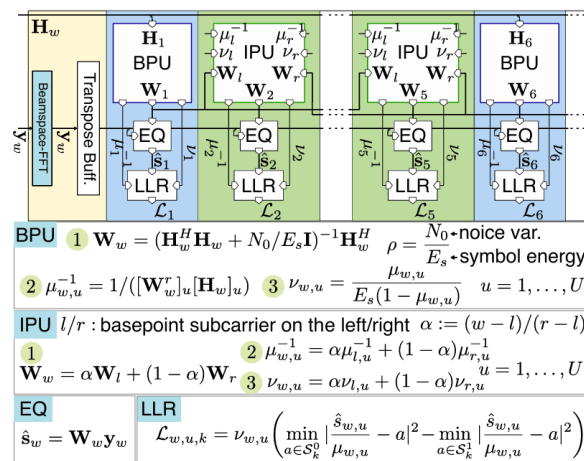
측정 칩은 Intel16 공정에서 3.17 mm²의 core area로 제작되었으며 16개의 tile DSP와 hub DSP를 포함해 256×8 uplink를 지원한다. 0.9 V, 550 MHz에서 16QAM detection throughput 17.6 Gb/s를 달성하고 687.6 mW를 소비하여 39.1 pJ/b의 energy efficiency를 보였다. 8명의 user에 대한 channel estimation latency는 461 ns이며, 기존 beamspace ASIC 대비 더 많은 antenna를 지원하면서 weight calculation까지 통합한다. 저자들이 비교한 최근 MIMO detector 대비 detection throughput은 최대 9.76배 높고, abstract 기준으로 area efficiency는 최대 6배 높으며 training latency는 1.78배 낮다. 이 결과는 2D beamspace와 tiled implementation이 대규모 array의 routing/scaling 문제까지 함께 다룬다는 점에서 의미가 있다.

#27-2 본 논문은 massive MU-MIMO-OFDM uplink에서 여러 subcarrier를 동시에 처리하는 soft-output detector를 제안한다. OFDM에서는 수천 개의 subcarrier마다 equalization을 수행해야 하므로 각 subcarrier에 완전한 preprocessing hardware를 복제하면 area와 power가 커지고, 반대로 serial processing을 택하면 latency와 buffering이 증가한다. 저자들은 이 문제를 해결하기 위해 channel impulse response가 소수 dominant tap에 집중된다는 frequency-domain correlation과, mmWave channel이 beamspace에서 sparse하다는 spatial property를 동시에 이용한다. 즉, 모든 subcarrier에서 LMMSE equalization matrix를 독립적으로 계산하지 않고 일부 basepoint에서만 정확한 preprocessing을 수행한 뒤 나머지는 interpolation으로 생성한다.



[그림 3] massive MU-MIMO-OFDM uplink와 frequency-domain correlation/beamspace sparsity 활용 개념

그림 3은 해당 논문이 동시에 활용하는 두 종류의 redundancy를 정리한다. 첫째, channel impulse response가 소수 dominant tap에 집중되기 때문에 인접 OFDM subcarrier의 channel과 equalization matrix가 높은 frequency-domain correlation을 갖는다. 따라서 모든 subcarrier에서 LMMSE preprocessing을 반복하는 대신 일정 stride마다 basepoint만 직접 계산하고 중간 subcarrier의 parameter는 interpolation으로 생성할 수 있다. 둘째, mmWave channel은 beamspace에서 sparse하므로 antenna-domain received signal을 spatial FFT로 변환하면 signal과 equalization matrix의 energy가 소수 beam에 집중된다. 저자들은 첫 번째 특성을 이용해 preprocessing hardware의 복제 비용을 줄이고, 두 번째 특성을 이용해 실제 MAC switching activity를 줄이는 방식으로 area와 power를 각각 다른 경로에서 절감한다.



[그림 4] 16-subcarrier soft-output detector top-level architecture와 BPU/IPU/EQ/LLR dataflow

그림 4의 fabricated ASIC은 16개의 parallel soft-output detector core를 두어 16개 subcarrier를 동시에 처리한다. 예시에서는 subcarrier 1, 6, 11, 16이 basepoint이며, 각

basepoint의 BPU가 LMMSE equalization matrix W 와 LLR 계산에 필요한 inverse bias term 및 post-equalization SINR을 생성한다. 그 사이 non-basepoint subcarrier는 IPU가 좌우 basepoint의 parameter를 위치에 따라 linear interpolation한다. 이후 EQ가 symbol estimate를 계산하고 LLR unit이 max-log approximation 기반의 soft output을 생성한다. 즉 expensive matrix preprocessing은 일부 basepoint에만 집중시키면서도 모든 subcarrier에서 soft-output detection을 유지하도록 architecture가 구성되어 있으며, IPU가 matrix를 column-by-column으로 처리하기 때문에 interpolated subcarrier를 위한 full matrix storage도 피할 수 있다.

BPU의 중심은 $U \times (U+1)/2$ 개의 PE로 구성된 triangular systolic array이며 $U=8$ 이므로 36개의 PE가 Gram computation, Cholesky decomposition, forward substitution, backward substitution의 네 phase를 순차적으로 수행한다. Phase마다 별도 accelerator를 두지 않고 동일 PE에 여러 data port와 multiplexer를 배치하여 hardware를 재사용하며 controller가 dataflow를 전환한다. Beamspace sparsity는 EQ와 BPU 내부의 signal silencing으로 실제 dynamic power 절감에 연결된다. 두 input magnitude가 threshold 이하인 complex MAC에서는 register update를 막아 연산을 skip하고, beamspace 신호의 증가된 fixed-point bit-width 문제는 variable-point(VP) number format과 FXP2VP/VP2FXP conversion을 통해 완화한다. VP 적용만으로 fully fixed-point design 대비 약 8%의 area가 줄어든다.

Implementation은 GlobalFoundries 22FDX에서 이루어졌고 core cell area는 1.3 mm^2 이다. 64×8 MIMO, QPSK부터 256-QAM까지 지원하며 0.8 V에서 286 mW를 소비하면서 7.3 Gb/s throughput과 39 pJ/b의 energy efficiency를 달성한다. Nominal 0.8 V에서 최대 clock frequency는 약 755 MHz이고, forward body bias를 사용하면 900 MHz 이상으로 상승해 8.8 Gb/s를 넘는 throughput을 얻을 수 있으나 leakage power가 증가하는 trade-off가 있다. Table 1에서 preprocessing을 포함하는 비교 대상들 중에서는 throughput과 energy efficiency가 가장 우수한 수준으로 보고된다.

정확도 측면에서 linear interpolation은 단순한 근사로 끝나지 않는다. 측정된 coded BLER는 antenna-domain(AD) mode와 beamspace-domain(BD) mode 모두 floating-point brute-force baseline과 대부분 유사하며, 가장 어려운 non-LoS 256-QAM 조건에서만 5G target인 10% BLER 지점에서 약 1 dB의 손실이 나타난다. Power 측정에서는 모든 subcarrier를 basepoint로 처리하는 brute-force baseline 대비 interpolation이 2.1배 power reduction과 3배 area reduction을 제공하고, beamspace sparsity를 이용한 silencing이 추가로 최대 1.4배 power reduction을 제공하여 전체적으로 약 3배의 power 절감 효과를 만든다. 즉, frequency-domain redundancy는 preprocessing 복제 비용을 줄이고 spatial sparsity는 실제 MAC switching activity를 줄이는 서로 다른 역할을 수행한다.

저자정보



이재성 박사과정 대학원생

- 소속 : 서울대학교
- 연구분야 : SNN 및 PIM
- 이메일 : jslee0122@snu.ac.kr
- 홈페이지 : <https://sites.google.com/view/ic3-snu>