

2026 VLSI Review

KAIST 인공지능반도체대학원 박사과정 하상우

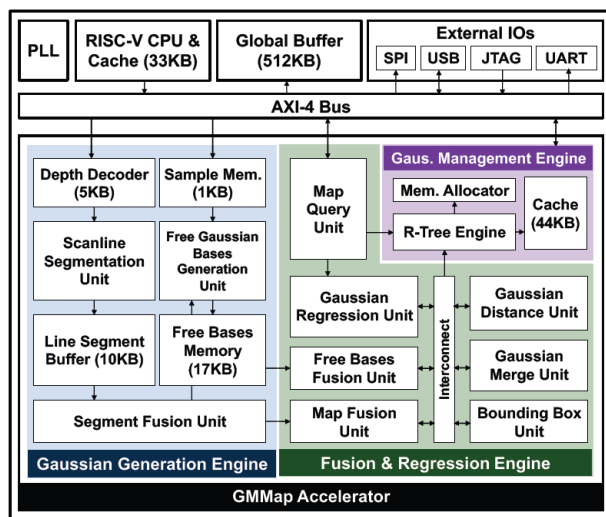
Session 3 Accelerators: Spatial and Immersive Visual Computing

이번 2026 IEEE/JSAP Symposium on VLSI Technology & Circuits의 Session C3에서는 spatial computing 및 immersive visual computing을 위한 accelerator SoC를 주제로 총 5편의 논문이 발표되었다. 최근 AR/VR, 로봇, embodied agent와 같은 edge device에서 3D 공간을 실시간으로 인지하고 재구성하려는 수요가 커지면서, 3D Gaussian 기반 표현, neural video codec, vision-language navigation, event 기반 센싱과 같은 새로운 workload가 등장하고 있다. 이들 workload는 연산량 뿐 아니라 불규칙한 자료구조 접근, sorting, entropy coding과 같은 non-GEMM 연산이 병목이 되는 경우가 많아, 알고리즘 특성을 회로 구조에 직접 반영하는 algorithm-hardware co-design이 핵심 설계 이슈로 부각되고 있다. 본 리뷰에서는 세션에서 발표된 다섯 편의 논문 #C3-1부터 #C3-5까지를 차례로 살펴보고자 한다.

#C3-1 본 논문은 MIT 연구진이 발표한 것으로, Gaussian 기반 3D occupancy map인 GMMap을 위한 6 mW급 SoC인 Gleanmer를 제안하였다. 3D occupancy map은 공간을 occupied, free, unexplored 영역으로 구분하여 안전한 자율 주행 및 AR/VR 상호작용을 가능하게 하지만, OctoMap과 같은 기존 방식은 하드웨어 가속을 적용하더라도 100 mW 이상을 소모하여 edge 전력 예산을 초과한다. GMMap은 occupied 및 free 영역을 3D Gaussian ellipsoid로 압축 표현하지만, 기존 embedded software 구현은 2 W 이상을 소모한다는 한계가 있었다.

이를 해결하기 위해 Gleanmer는 세 가지 co-optimization을 적용하였다. 첫째, 전체 map construction 에너지의 최대 65%를 차지하던 free Gaussian bases 생성을 수천 개의 line segment가 아닌 occupied Gaussian으로부터 직접 계산하여 construction 에너지를 22~63% 절감하였다. 둘째, trajectory 상의 인접한 query coordinate들이 유사한 R-Tree 경로를 순회한다는 점에 착안하여, 16개 coordinate를 하나의 bounding box로 묶는 batch query를 통해 query throughput을 4~10배 높이고 query 에너지를 74~81% 절감하였다. 셋째, scanline segmentation의 slope를 1-cycle로 근사하고 Gaussian precision을 32-bit에서 19-bit로 낮추는 approximate computing으로 accelerator 면적을 38% 줄였다.

16nm CMOS로 제작된 칩은 640×480 depth image를 88 fps 이상으로 처리하며, 6 mW 이하의 전력으로 NVIDIA Jetson TX2 대비 341배, OctoMap 가속기인 OMU 대비 44배 낮은 전력 소모를 달성하였다.

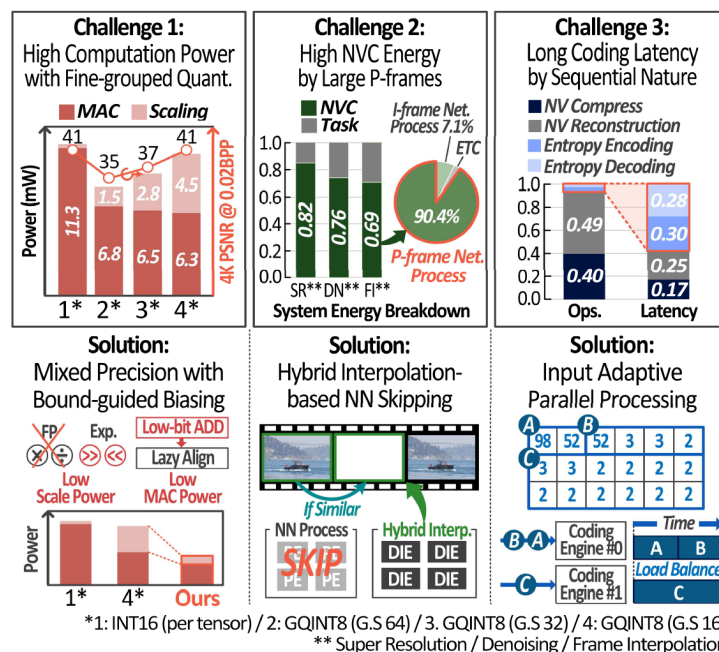


[그림 1] Gleanmer SoC의 System architecture

#C3-2 본 논문은 KAIST 연구진이 발표한 NOVA로, 단일 칩에서 decoding-processing-encoding을 end-to-end로 수행하는 최초의 fully integrated neural video SoC를 제안하였다. 기존 neural video processor들은 super resolution, denoising과 같은 video task 가속에만 집중하여 압축 및 전송을 위한 별도의 codec 하드웨어가 필요했고, implicit NVC는 video별 학습에 서버급 GPU로 1~2일이 소요되어 decode-only에 머물렀다. 단일 forward pass로 인코딩이 가능한 explicit NVC의 등장으로 단일 칩 통합의 가능성이 열렸지만, quantization 민감성, P-frame 처리가 시스템 에너지의 69~82%를 차지하는 문제, symbol 단위 순차 연산으로 인한 entropy coding latency가 새로운 병목으로 부각되었다.

NOVA는 이를 세 가지 기법으로 해결하였다. 첫째, bound-guided mixed quantization(BMQ)은 inlier를 INT4로 표현하면서 shared exponent bias를 shift 연산으로 대체하여 FP scaling 전력을 91.9% 줄이고, shift-ordered adder tree를 갖춘 bound-guided unified(BU) core와 함께 연산 전력을 41.7% 절감하였다. 둘째, temporal interpolation unit(TIU)은 NVC 초기 layer의 flow 정보를 활용해 temporally predictable한 frame의 neural processing을 선택적으로 생략하고, tiered DMA를 통해 TIU와 core의 병렬 동작을 지원하여 encoding/decoding 에너지를 각각 71.2%/71.8% 절감하였다. 셋째, adaptive entropy coding unit(AECU)은 entropy 분포에 따라 chunk 경계를 동적으로 결정하고 load balancing controller로 병렬 engine에 분배하여 hardware utilization을 91.6%까지 높

이고 coding latency를 최대 79.2% 줄였다. 28nm CMOS로 제작된 칩은 4K UVG benchmark에서 31.4 TOPS/W의 에너지 효율을 달성하였으며, encoding은 5.3 mJ/frame과 13.7 ms/frame으로 edge GPU 대비 76.7배 빠르고, decoding은 6.3 mJ/frame으로 4K 70.8 FPS를 달성하였다.

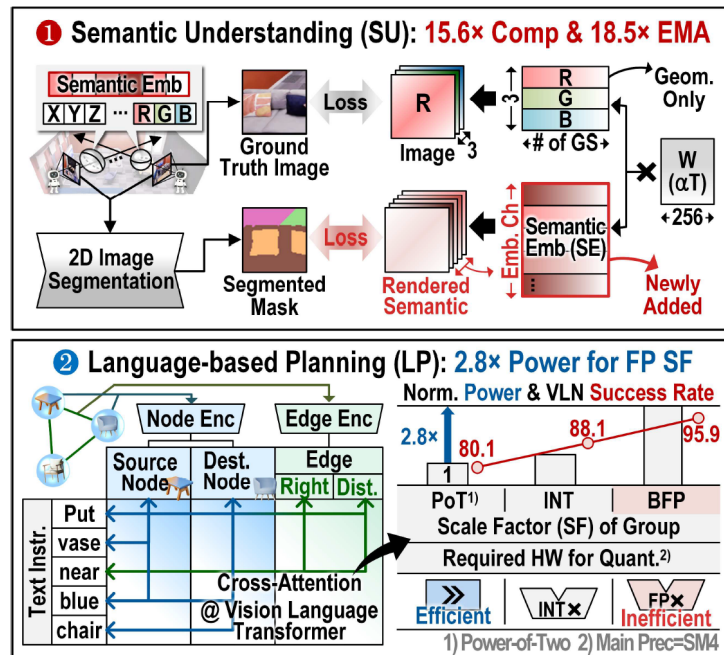


[그림 2] Video Codec HW의 문제점과 논문 #C3-2의 해결법

#C3-3 본 논문 역시 KAIST 연구진이 발표한 것으로, embodied agent를 위한 spatial reasoning 기반 vision language navigation(VLN) 가속기인 SR-VLNA를 제안하였다. 기존 point planning 가속기들은 미리 정의된 목표 좌표를 향한 occupancy 기반 경로 계획만 수행하여 동적인 주변 환경과 언어 명령을 이해하지 못한다. 제안된 VLN 시스템은 3D Gaussian splatting 기반 semantic understanding(SU)과 language-based planning(LP)으로 구성되지만, edge GPU에서는 semantic embedding 추가로 인한 15.6배의 연산량과 18.5 배의 external memory access, cross-attention의 넓은 activation 범위로 인한 FP group scaling 전력, SU와 LP의 상이한 sparsity 패턴이 병목이 된다.

SR-VLNA는 네 가지 기법을 제안하였다. Embedding skipping unit(ESU)은 가려진 객체의 semantic embedding과 중요도가 낮은 channel을 제거하여 연산량을 65.6%, external memory access를 77.7% 줄였고, mapping region decision unit(MRDU)은 balanced KD-tree 기반 bounding box 추출로 정적 객체의 Gaussian redundancy를 제거하여 SU latency를 93.4% 절감하였다. Context-aware group quantization unit(CAGQU)은

instruction 유형과 edge 수에 따라 power-of-two와 INT/FP scale factor를 선택적으로 사용하여 FP SF 사용을 92.9% 줄였으며, heterogeneous sparsity-aware adder tree(HSAT)는 sparse한 데이터에는 carry-suppressed adder tree를, dense한 데이터에는 zero count adder tree를 할당하여 에너지 효율을 2.3~2.7배 개선하였다. 28nm FDSOI로 제작된 칩은 SU에서 0.004~0.021 $\mu\text{J}/\text{point}$ 의 state-of-the-art 성능을 보였고, LP를 최초로 시연하여 0.25~0.96 mJ/task를 달성하였으며, 시스템 레벨에서 5.0~23.9 mJ/meter로 edge GPU 대비 압도적으로 낮은 에너지를 기록하였다.



[그림 3] Spatial reasoning 기반 VLN의 문제점

#C3-5 본 논문은 Tsinghua University 연구진이 발표한 것으로, 3D/4D Gaussian splatting 실시간 렌더링을 위한 12nm spatial video rendering processor인 Uni4D를 제안하였다. 4DGS는 Hex-plane encoding과 MLP 기반 deformation으로 3DGS를 동적 장면으로 확장 하지만, 하드웨어 관점에서는 K-plane의 3D indexing에 필요한 대형 2D register array, tile 기반 preprocessing에서의 순차적인 Gaussian box-bounding, software 방식 radix sorting으로 인한 frame당 수백 MB의 external memory access가 병목이 된다.

Uni4D는 K-plane sampling과 MLP deformation을 단일 datapath로 통합한 UniMAC deformation engine과, 64×64 grid를 16×16 sub-grid로 분할하는 K-plane weights segmentation(KWS)을 통해 K-plane buffer 면적을 84% 절감하였다. Non-blocking KVP generation engine은 boundary tile filtering unit으로 tile 범위를 shift와 decoder만으로 도출하고 다수의 key-value generation unit에 병렬 분배하여 preprocessing 시간을 74% 단

축하였으며, hierarchical radix sorting engine(HRSE)은 global tile sorting과 on-chip depth sorting을 분리하고 4-bit partial sub-sorting과 rescheduled parallel sorting workload를 적용하여 external memory access를 frame당 464 MB에서 126 MB로 줄였다. 12nm CMOS로 제작된 5.29 mm² 칩은 0.16~0.52 mJ/frame의 에너지로 HyperNeRF 시퀀스를 74~76 fps, synthetic D-NeRF 장면을 최대 232 fps로 렌더링하며, 기존 NeRF/3DGS 프로세서 대비 한 자릿수 이상 낮은 frame당 에너지를 달성하였다.

종합적으로 Session C3는 3D Gaussian이 occupancy mapping(#C3-1), semantic understanding(#C3-3), spatial video rendering(#C3-5)에 이르기까지 edge spatial computing의 공통 표현으로 자리잡고 있으며, 이를 뒷받침하는 전용 하드웨어가 mW급 전력 예산에서 실시간 성능을 달성하는 단계에 도달했음을 보여주었다. 또한 R-Tree 순회, entropy coding, KD-tree 탐색, radix sorting, graph 구성과 같이 GPU가 비효율적인 불규칙 연산을 전용 engine으로 가속하고, sparsity와 temporal locality 같은 workload 특성을 quantization 및 adder tree 수준까지 반영한 점이 세션 전반의 공통된 설계 철학으로 읽혔다.

저자정보



하상우 박사과정 대학원생

- 소속 : KAIST
- 연구분야 : 딥러닝 가속기 설계
- 이메일 : sangwoo_ha@kaist.ac.kr
- 홈페이지 : <https://ssl.kaist.ac.kr>