UE8M0-TPU Architecture v0.2 — Full Report

# 1) Kick-off Summary (v0.1 baseline)

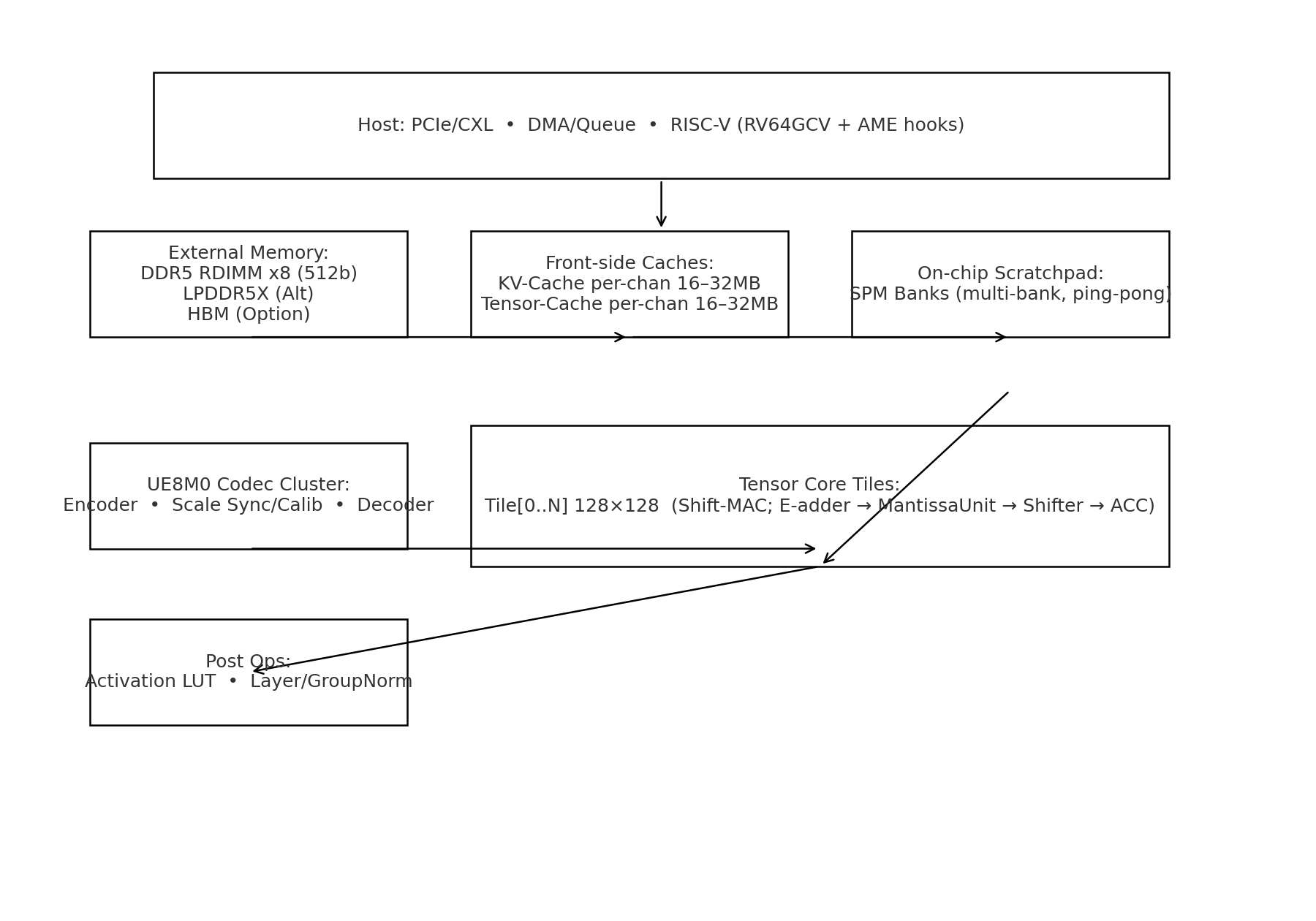
철학: 곱셈기 없는 2^E shift + FP8 mantissa → 대규모 연산을 저전력·고효율로.  
주요 블록: Host&DMA, External DRAM(HBM/LPDDR), On-chip Scratchpad, UE8M0 Codec Cluster, Tensor Core Tiles, Activation/Norm 후처리.  
데이터플로우: Conv=Weight-stationary, GEMM=Output-stationary.  
포맷: FP8 mantissa + UE8M0 exponent group header(32 elems).  
PE: E-adder → MantissaUnit(8×8 MUL/LUT) → Shifter → 24b ACC.  
리스크: MUL 비용, Scale saturation → LUT 근사/Residual 경로로 완화.

# 2) Architecture 수정분 (v0.2)

• 메인 클럭 정책: DDR SKU=0.50 GHz 고정, HBM SKU=0.50~0.55 GHz sweet spot.  
• DDR SKU는 Front-side Cache 증설(채널당 16–32 MB + 공유 eDRAM 256–512 MB)로 유효 대역폭 확장.  
• DVFS 단순화: DDR=0.50 GHz 유지, 필요시 0.60/0.70/0.80 GHz 단계 상승. HBM=0.50 GHz 기본, 연산 부족시 0.55 GHz.  
• Tensor Core Tiles, Codec Cluster, Scratchpad 구조는 v0.1과 동일.

# 3) Architecture Diagram (v0.2)

아래는 v0.2 기준 SoC Top 구조도입니다.

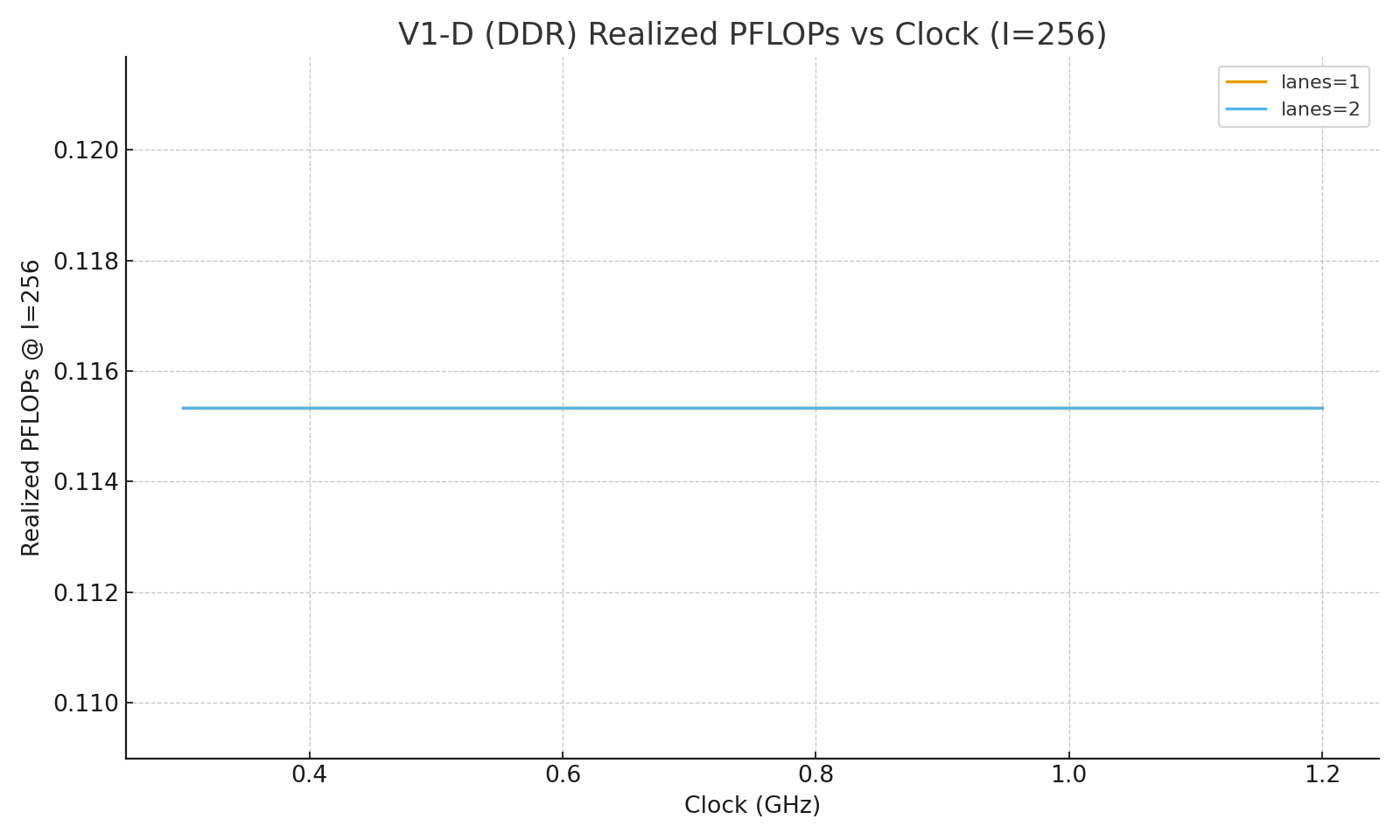


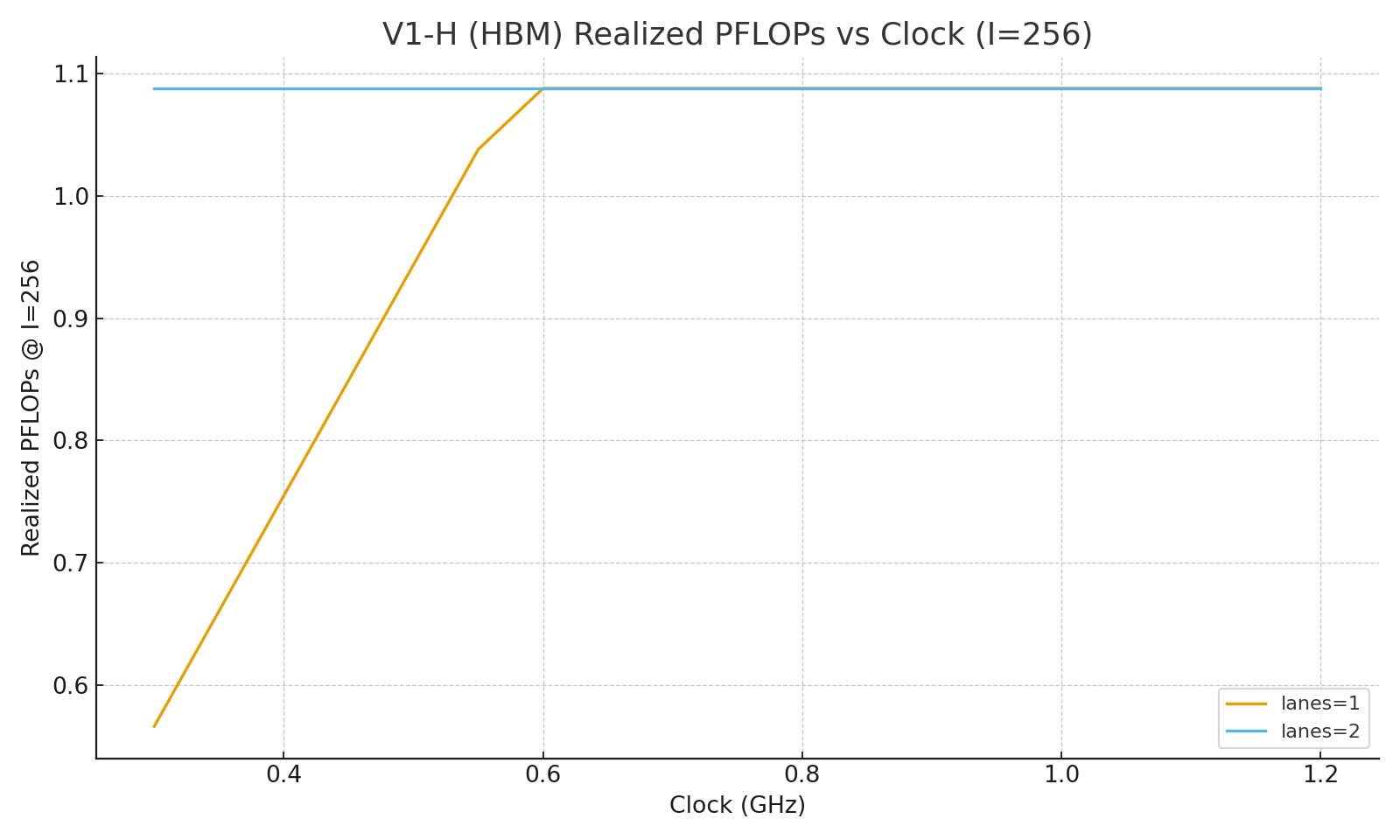
# 4) 성능 스윗스팟 요약표

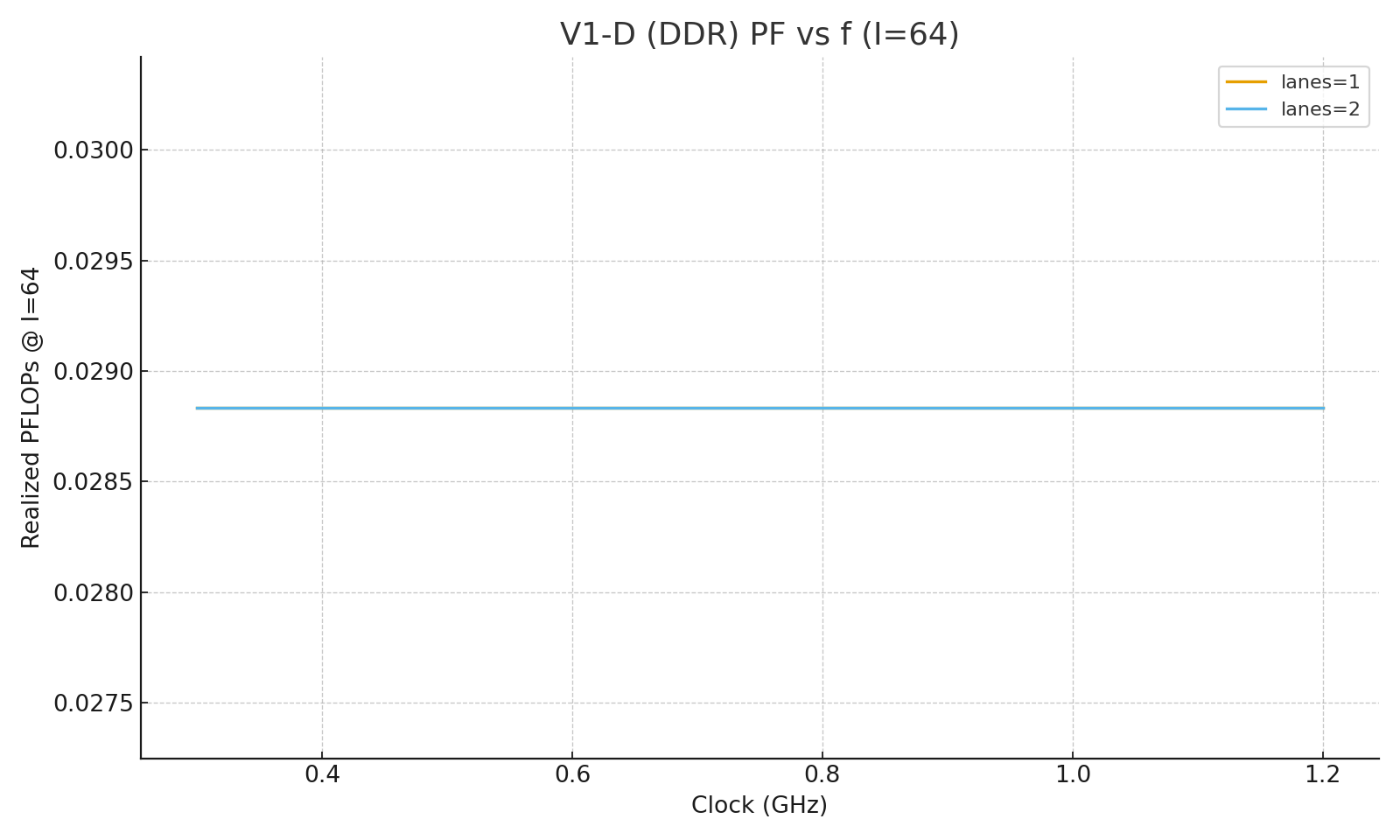
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SKU | 클럭(GHz) | lanes | I=256 Realized PF/s | Perf/W (TF/W, lin) |
| V1-D\_DDR | 0.5 | 1 | 0.115 | 0.51 |
| V1-D\_DDR | 0.5 | 2 | 0.115 | 0.51 |
| V1-H\_HBM | 0.5 | 1 | 0.944 | 2.18 |
| V1-H\_HBM | 0.5 | 2 | 1.088 | 2.51 |
| V1-H\_HBM | 0.55 | 1 | 1.038 | 2.28 |
| V1-H\_HBM | 0.55 | 2 | 1.088 | 2.39 |

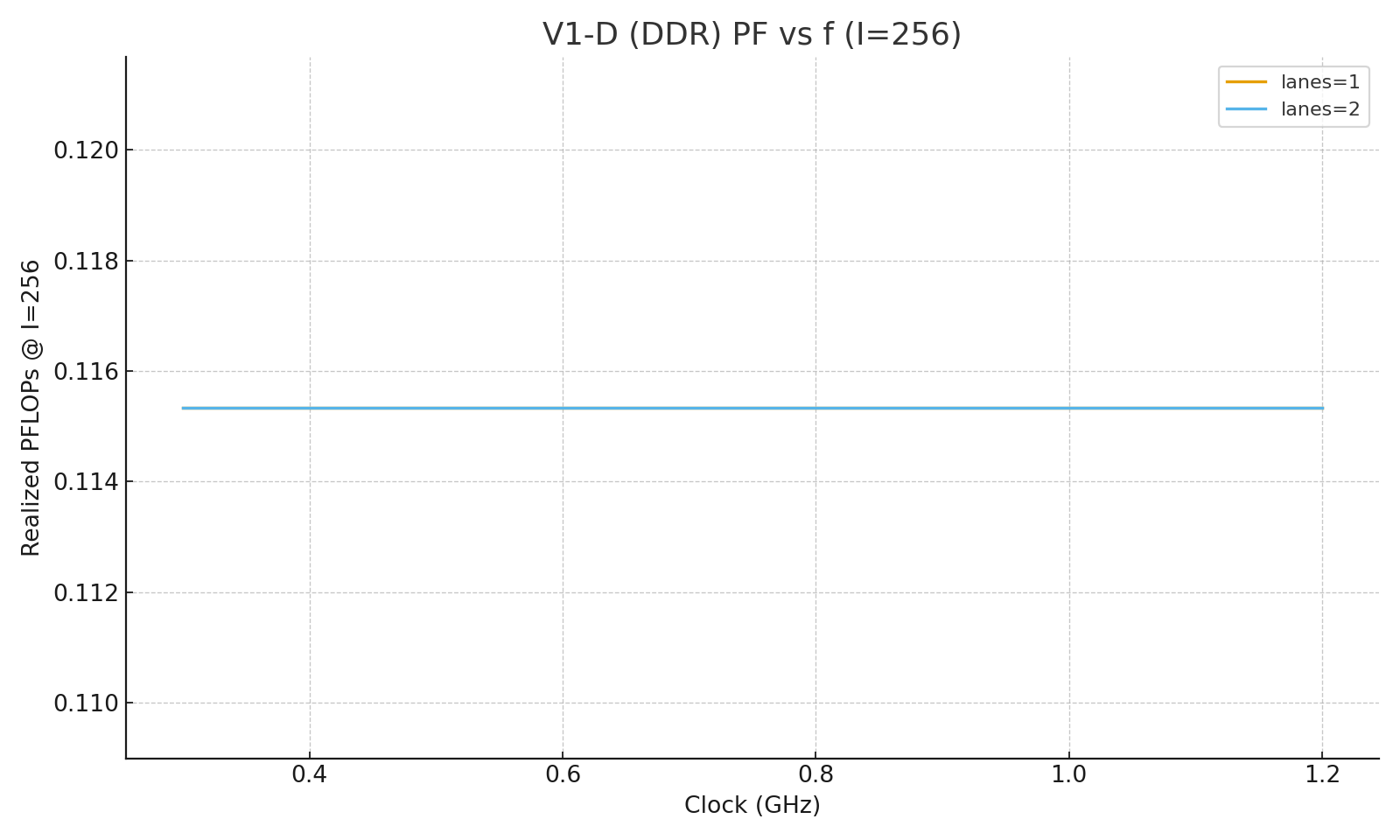
# 5) 성능 그래프

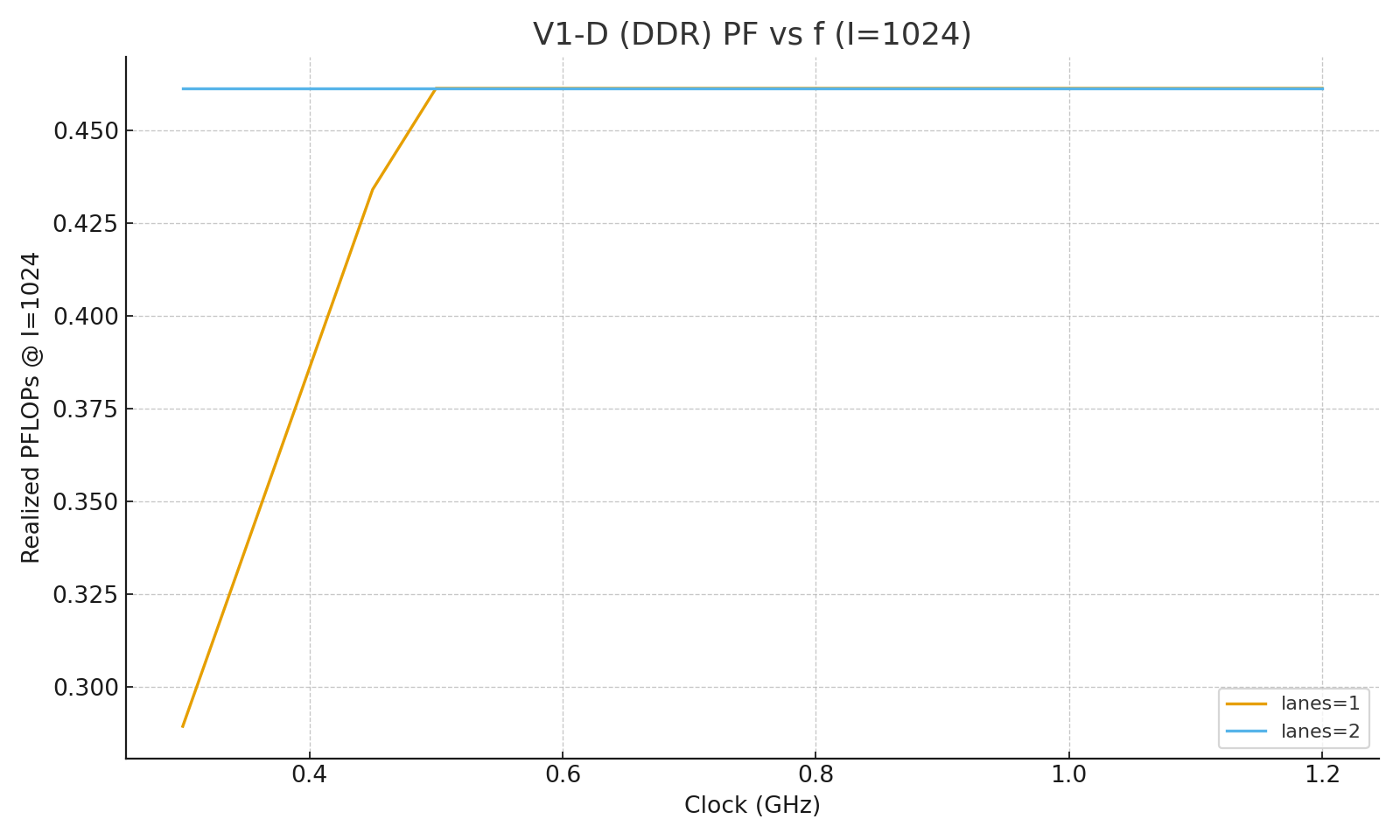
아래 그래프들은 I=256 기준 및 다양한 연산집약도에서의 클럭 대비 성능/효율을 보여줍니다.

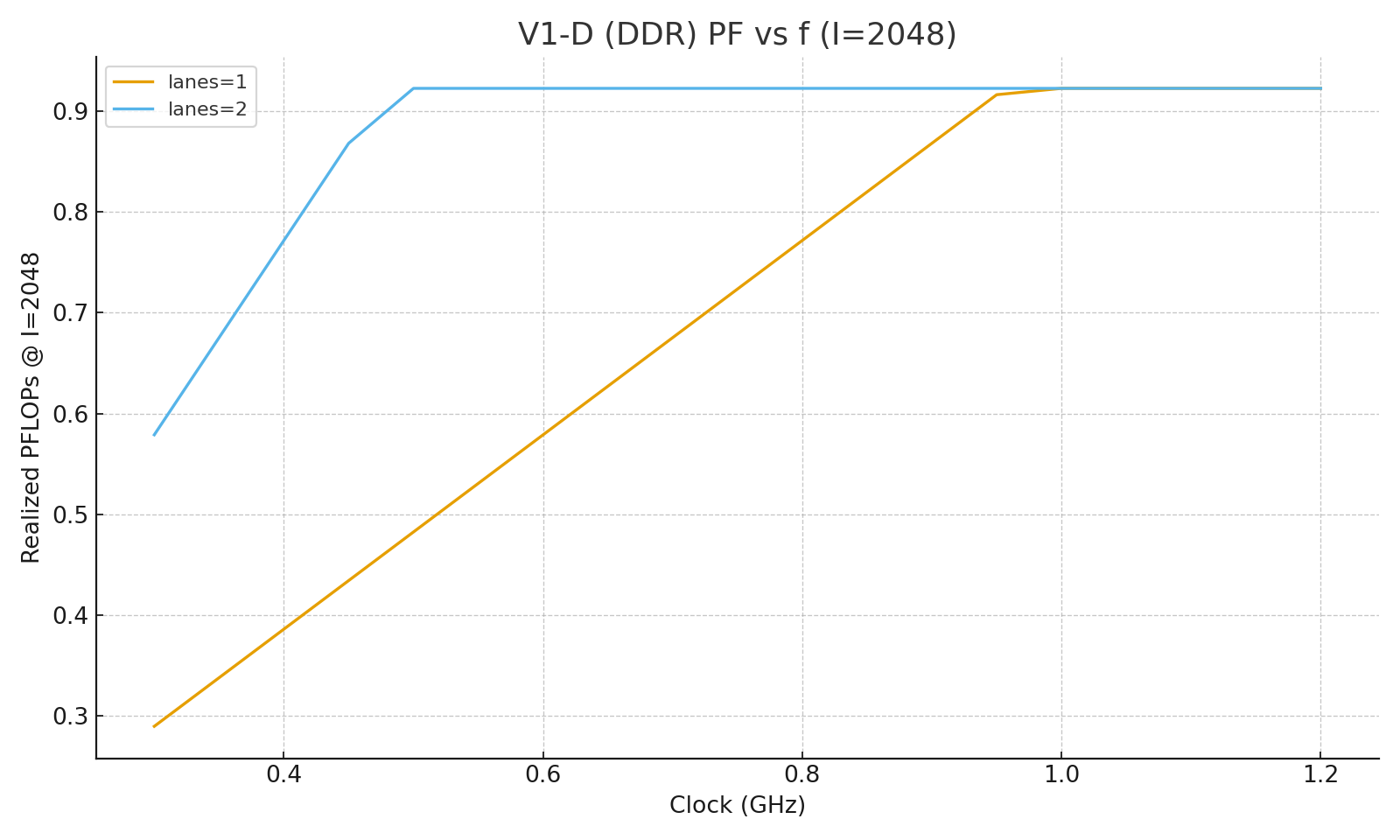


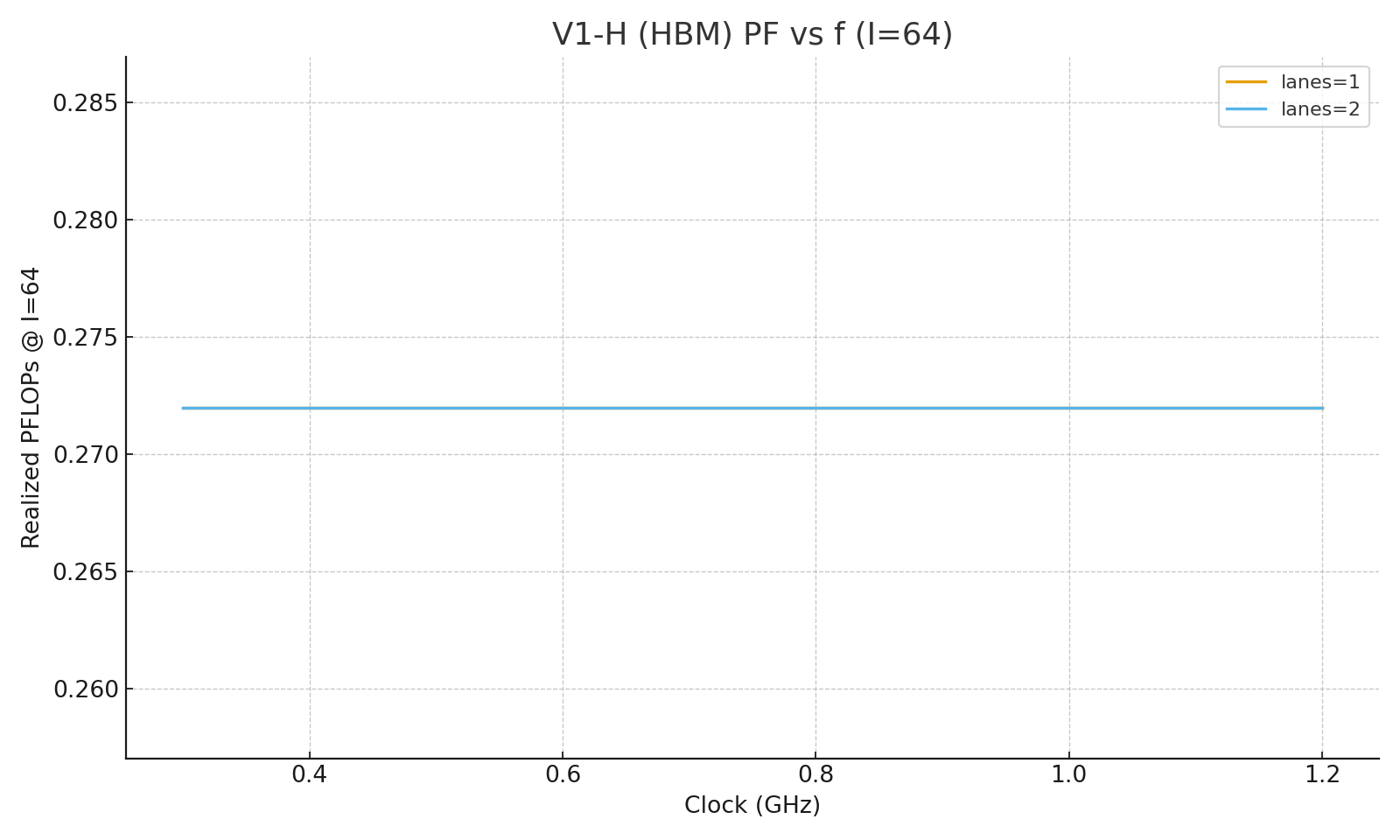


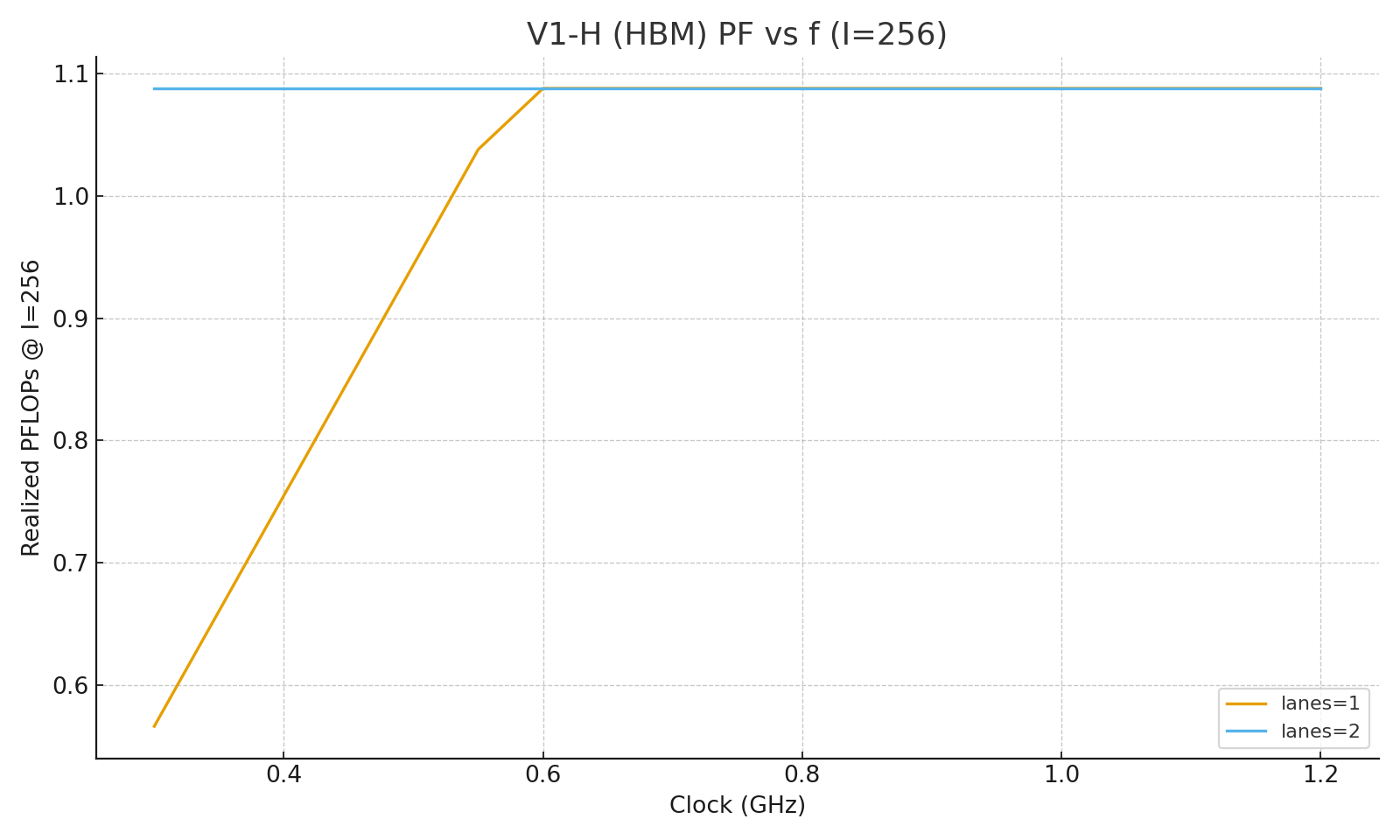


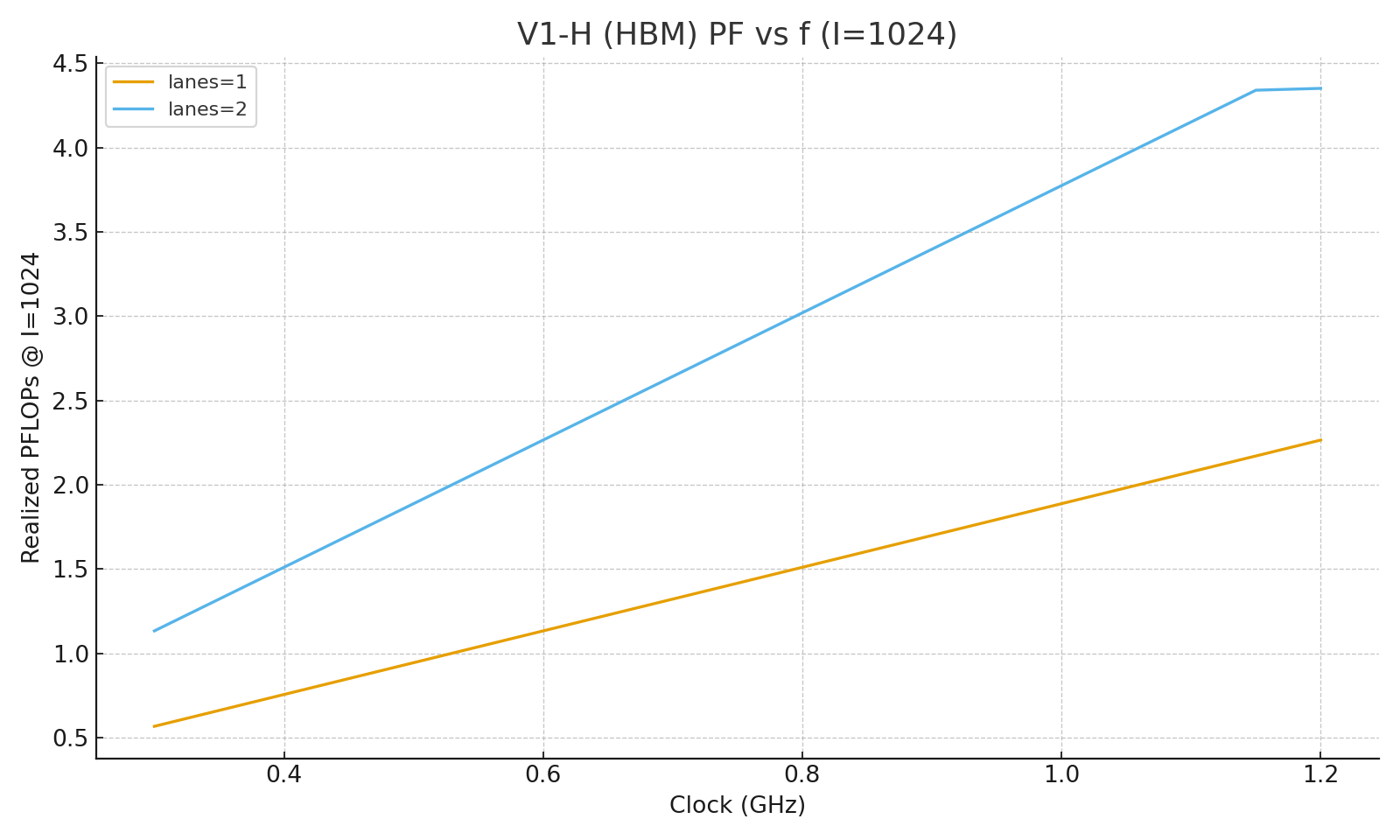


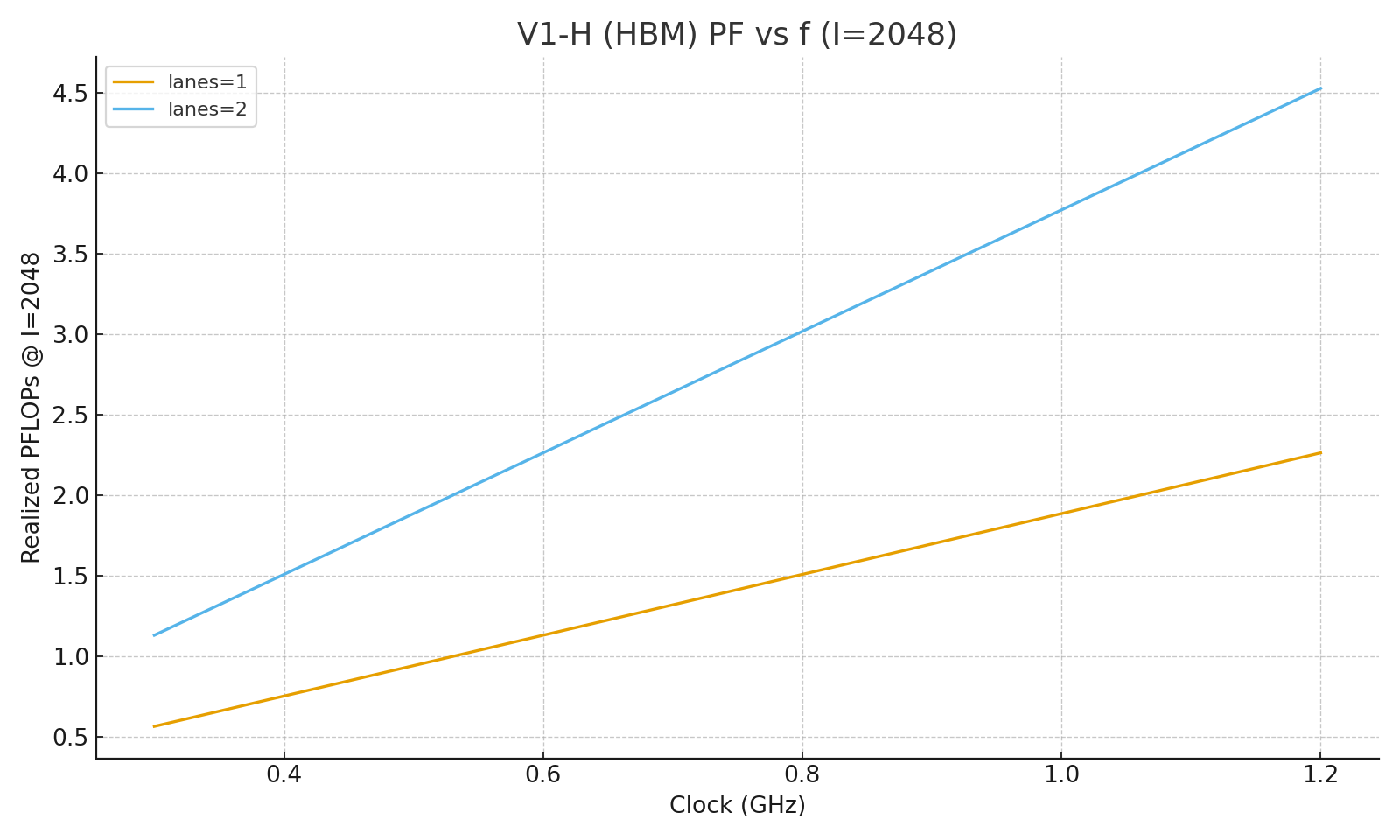












# 6) 요약

• DDR SKU: 0.50 GHz 고정, 캐시 증설 시 성능 개선.  
• HBM SKU: 0.50~0.55 GHz sweet spot, lanes=1일 때 0.55 GHz에서 roofline 도달.  
• 전체적으로 언더클럭 + 캐시 강화 전략이 Perf/W 최적화를 달성.