TF‑NPUv1 (UE8M0) Design Batch — Clock-under-DRAM & DVFS Policy v0.2

요약: ASIC-miner식 언더클럭/언더볼트 전략을 TF‑NPUv1에 적용. DDR SKU는 완전 메모리 바운드, HBM SKU는 0.55 GHz(lanes=1) 또는 0.30 GHz(lanes=2) 부근에서 Perf/W 최적.

# 1) 메서드

- 128×128 PEs/tile, 1 MAC/clk/PE(2 FLOPs), 효율: DDR 0.92 / HBM 0.90

- DDR5 8ch×64b @8.8GT/s → 450.6 GB/s 유효, HBM3e → 4.25 TB/s 유효

- 연산집약도 I ∈ {32..4096} FLOPs/Byte, f ∈ [0.30, 1.20] GHz sweep

- 전력모델: P = P\_static(30%) + P\_dyn·(f/f0)^exp, exp∈{1,2}

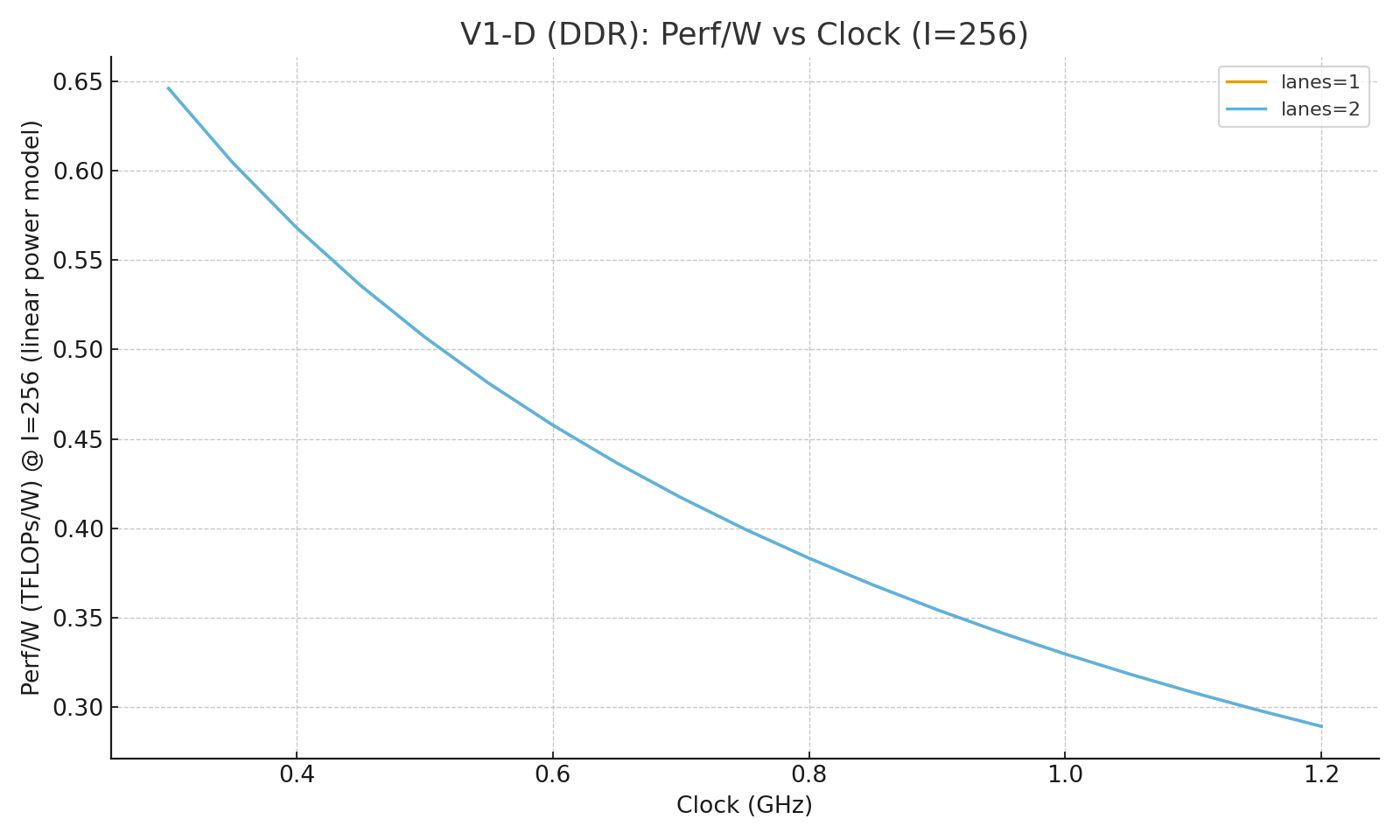
# 2) 결과 하이라이트

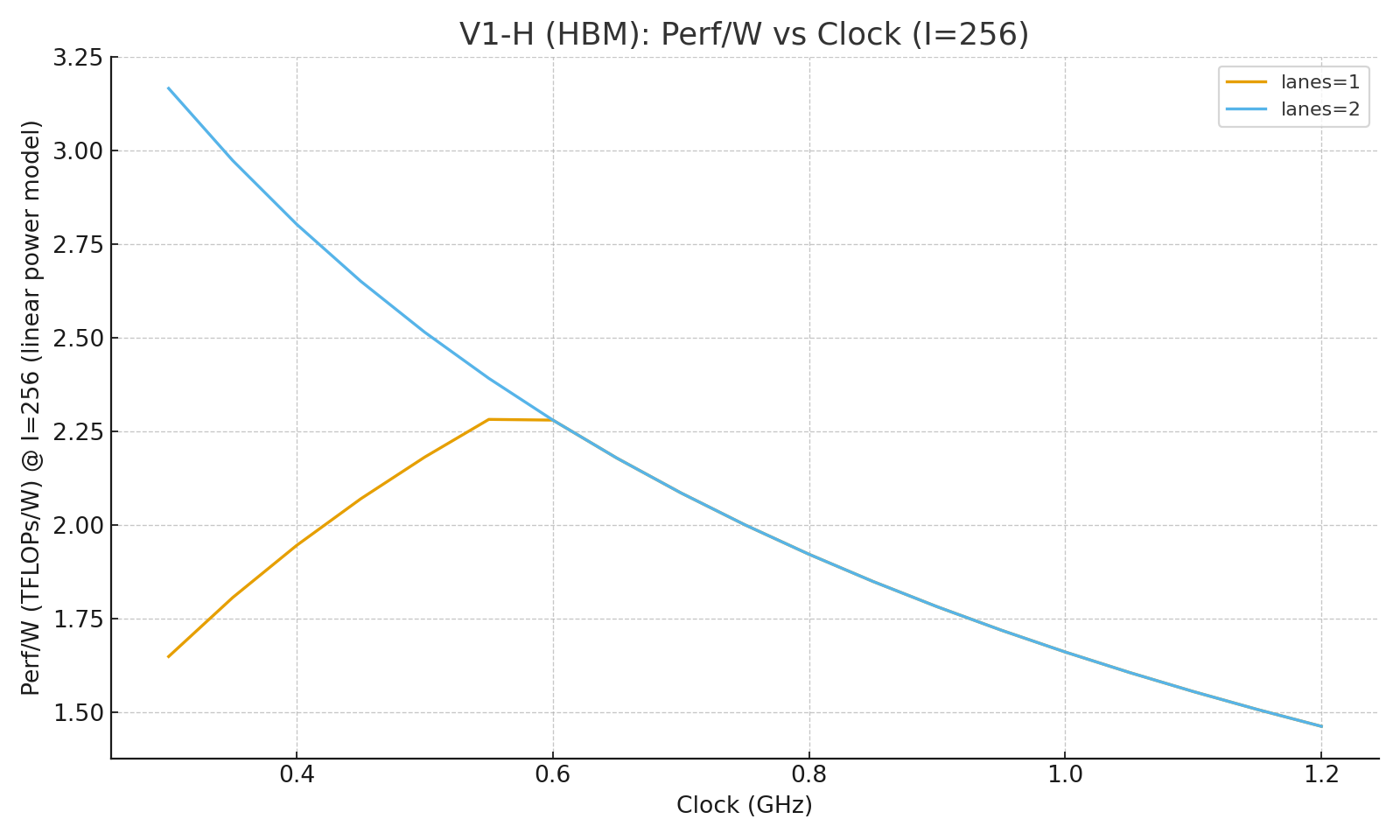
• V1‑D(DDR): I=256 기준 f를 0.30 GHz까지 낮춰도 PF/s ≈ 0.115로 평탄(완전 메모리 바운드).

• V1‑H(HBM): lanes=1은 f≈0.55 GHz에서 지붕선 도달, lanes=2는 0.30 GHz에서도 지붕선 도달 → 저클럭이 최적.

• 권고: 워크로드 실시간 I 추정 → f를 지붕선 직전으로 맞추는 DVFS(로컬 µ컨트롤러) 적용.

# 3) Perf/W vs Clock (I=256)





# 4) DVFS 정책표 (요약)

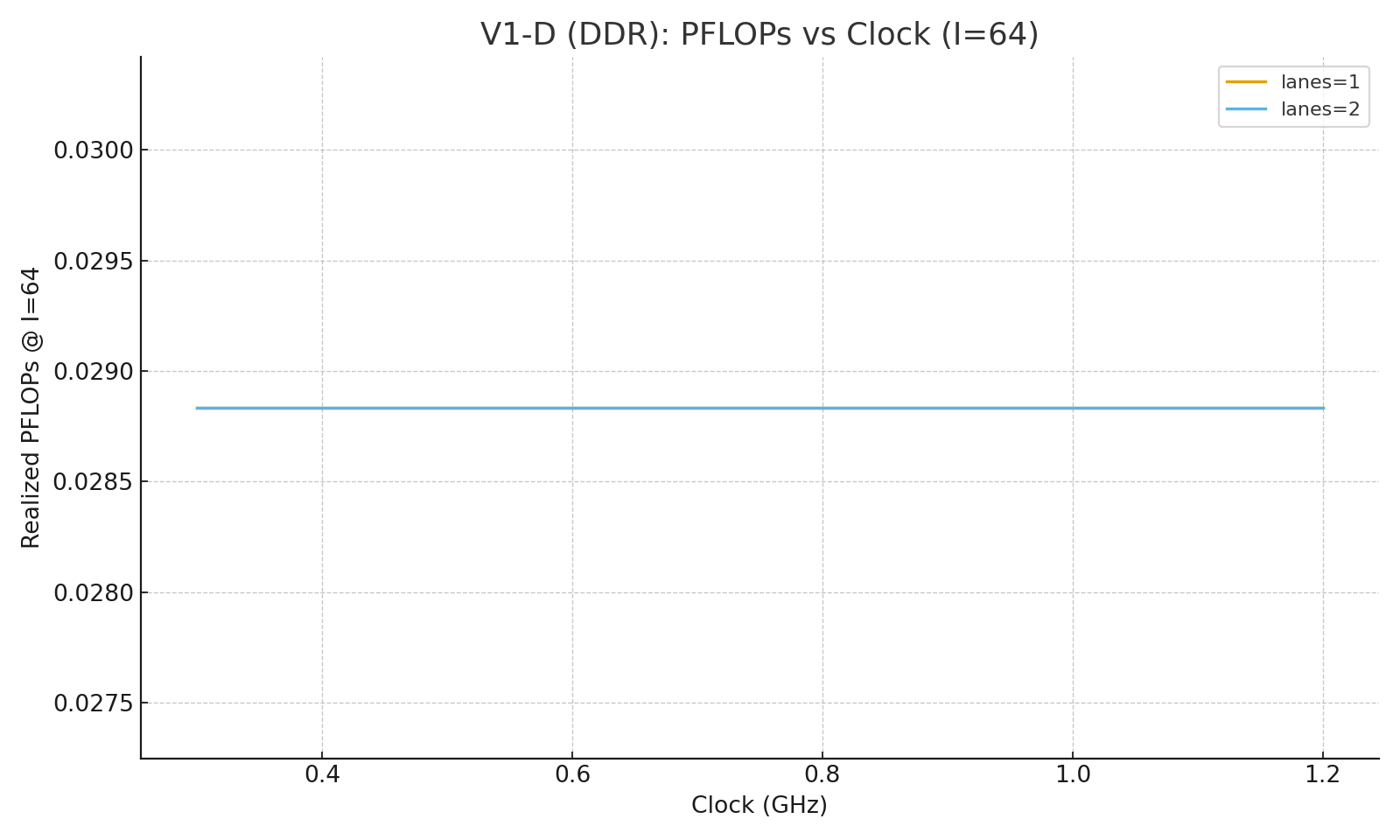
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SKU | lanes | I(FLOPs/B) | f\_target(GHz) | Realized PF | Power\_lin(W) |
| V1-D\_DDR | 1 | 32 | 0.30 | 0.014 | 178.5 |
| V1-D\_DDR | 1 | 64 | 0.30 | 0.029 | 178.5 |
| V1-D\_DDR | 1 | 128 | 0.30 | 0.058 | 178.5 |
| V1-D\_DDR | 1 | 256 | 0.30 | 0.115 | 178.5 |
| V1-D\_DDR | 1 | 512 | 0.30 | 0.231 | 178.5 |
| V1-D\_DDR | 1 | 1024 | 0.50 | 0.461 | 227.5 |
| V1-D\_DDR | 1 | 2048 | 0.95 | 0.916 | 337.8 |
| V1-D\_DDR | 1 | 4096 | 0.30 | 0.289 | 178.5 |
| V1-D\_DDR | 2 | 32 | 0.30 | 0.014 | 178.5 |
| V1-D\_DDR | 2 | 64 | 0.30 | 0.029 | 178.5 |
| V1-D\_DDR | 2 | 128 | 0.30 | 0.058 | 178.5 |
| V1-D\_DDR | 2 | 256 | 0.30 | 0.115 | 178.5 |
| V1-D\_DDR | 2 | 512 | 0.30 | 0.231 | 178.5 |
| V1-D\_DDR | 2 | 1024 | 0.30 | 0.461 | 178.5 |
| V1-D\_DDR | 2 | 2048 | 0.50 | 0.923 | 227.5 |
| V1-D\_DDR | 2 | 4096 | 0.95 | 1.833 | 337.8 |
| V1-H\_HBM | 1 | 32 | 0.30 | 0.136 | 343.6 |
| V1-H\_HBM | 1 | 64 | 0.30 | 0.272 | 343.6 |
| V1-H\_HBM | 1 | 128 | 0.30 | 0.544 | 343.6 |
| V1-H\_HBM | 1 | 256 | 0.60 | 1.088 | 477.3 |
| V1-H\_HBM | 1 | 512 | 1.15 | 2.171 | 722.3 |
| V1-H\_HBM | 1 | 1024 | 0.30 | 0.566 | 343.6 |
| V1-H\_HBM | 1 | 2048 | 0.30 | 0.566 | 343.6 |
| V1-H\_HBM | 1 | 4096 | 0.30 | 0.566 | 343.6 |
| V1-H\_HBM | 2 | 32 | 0.30 | 0.136 | 343.6 |
| V1-H\_HBM | 2 | 64 | 0.30 | 0.272 | 343.6 |
| V1-H\_HBM | 2 | 128 | 0.30 | 0.544 | 343.6 |
| V1-H\_HBM | 2 | 256 | 0.30 | 1.088 | 343.6 |
| V1-H\_HBM | 2 | 512 | 0.60 | 2.176 | 477.3 |
| V1-H\_HBM | 2 | 1024 | 1.15 | 4.341 | 722.3 |
| V1-H\_HBM | 2 | 2048 | 0.30 | 1.132 | 343.6 |
| V1-H\_HBM | 2 | 4096 | 0.30 | 1.132 | 343.6 |

# 5) DVFS 의사코드

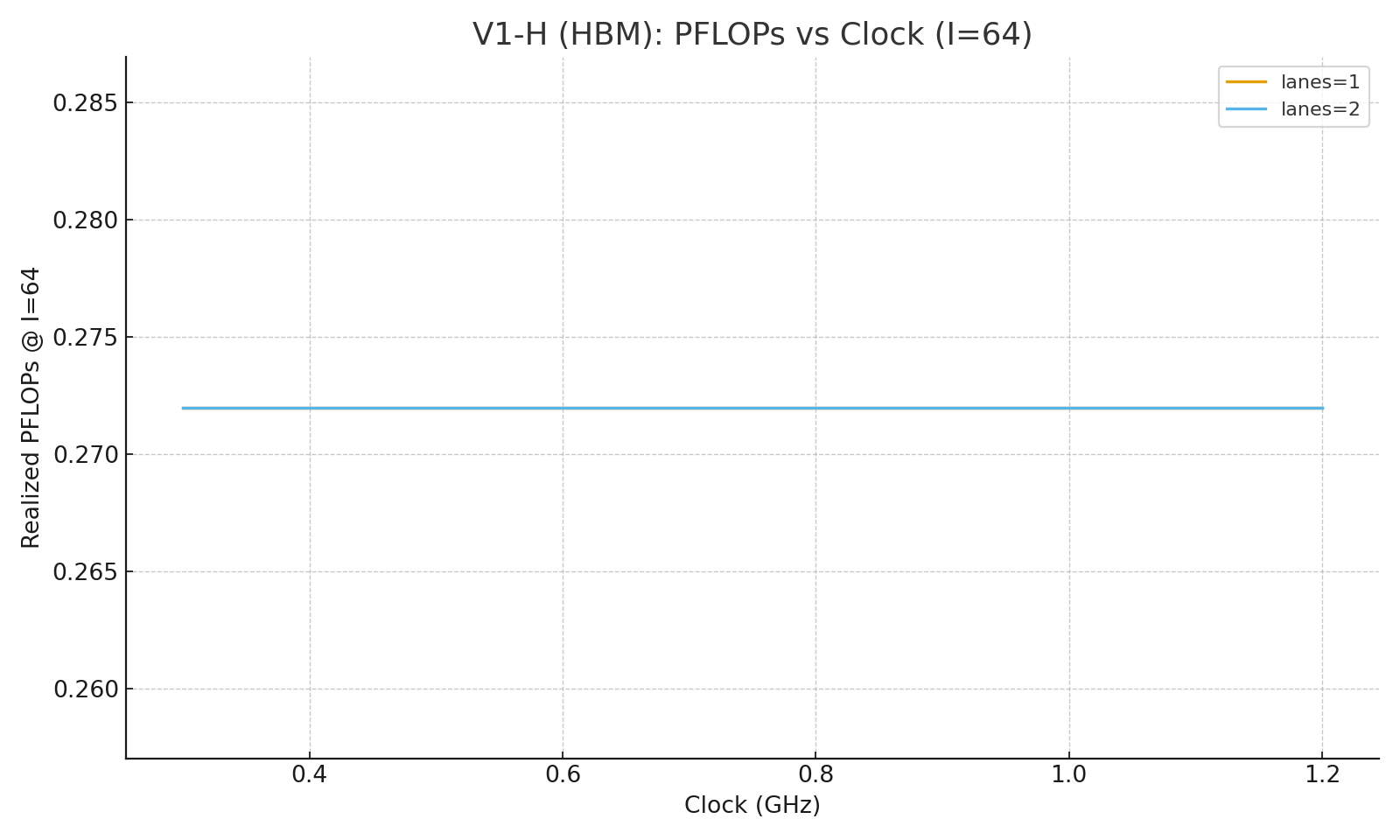
/\* Run-time DVFS controller (per-SKU, per-lanes)  
 \* Inputs: measured\_I (FLOPs/Byte), lanes, SKU, headroom=0.98  
 \*/  
float select\_freq\_GHz(float measured\_I, int lanes, SKU sku) {  
 // 1) map measured\_I to the nearest higher bin  
 int bins[] = {32,64,128,256,512,1024,2048,4096};  
 int Ibin = bins[sizeof(bins)/sizeof(bins[0]) - 1];  
 for (int i=0;i<sizeof(bins)/sizeof(bins[0]);++i) if (measured\_I <= bins[i]) { Ibin = bins[i]; break; }  
 // 2) lookup frequency target  
 float f = LUT[sku][lanes][Ibin];  
 // 3) apply guard rails (thermal, power budget)  
 return clamp(f, F\_MIN, F\_MAX);  
}

# 부록: PF vs f 곡선 (샘플)

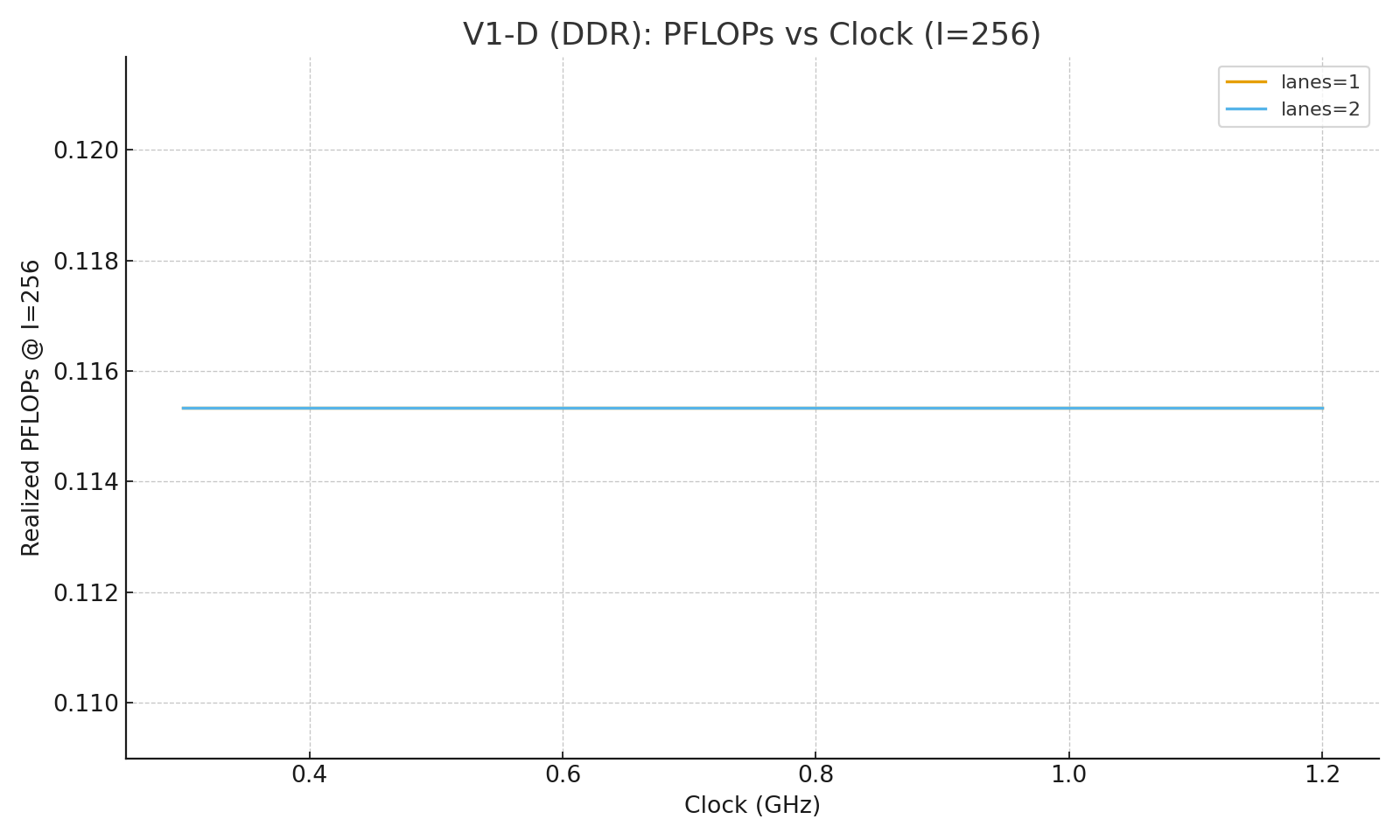
V1‑D (I=64)



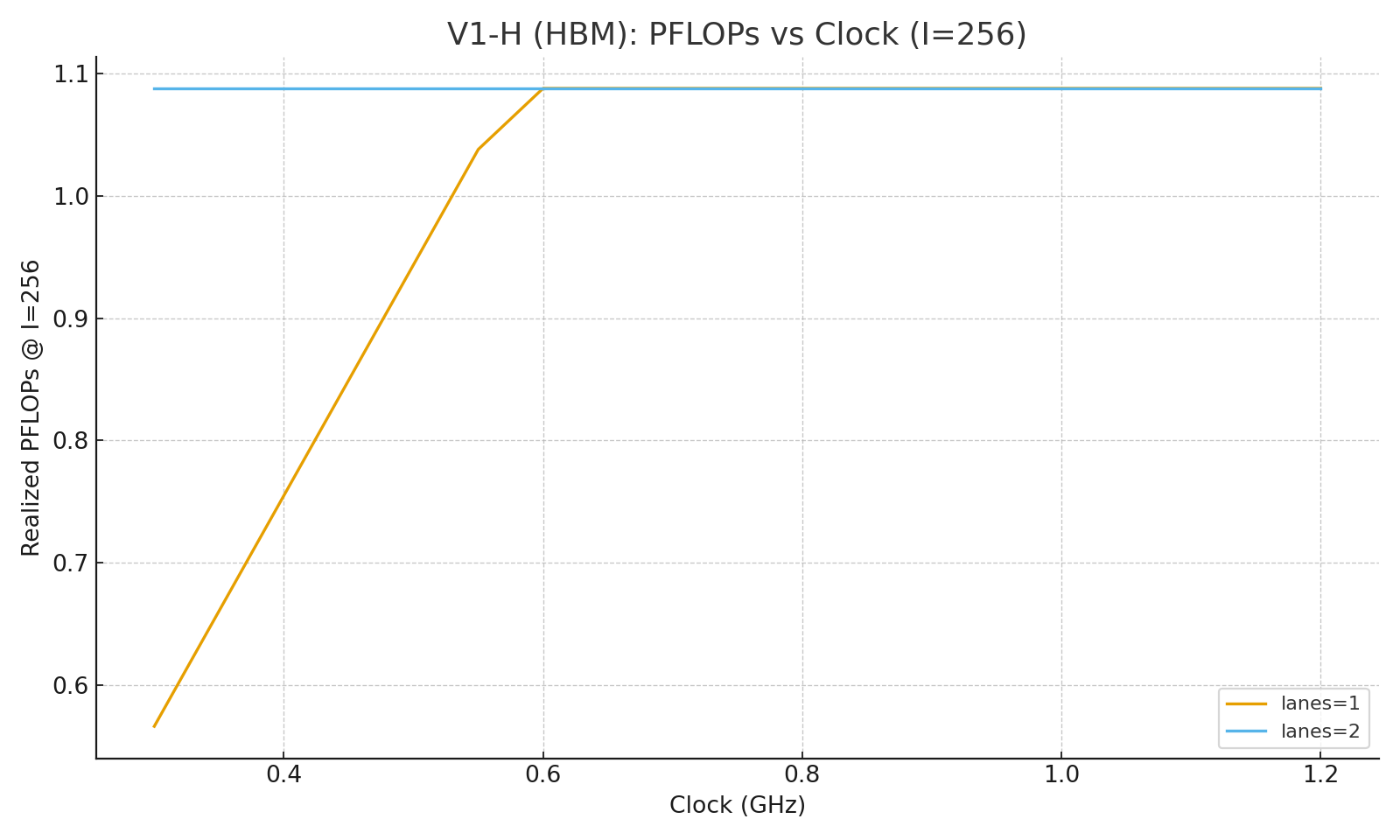
V1‑H (I=64)



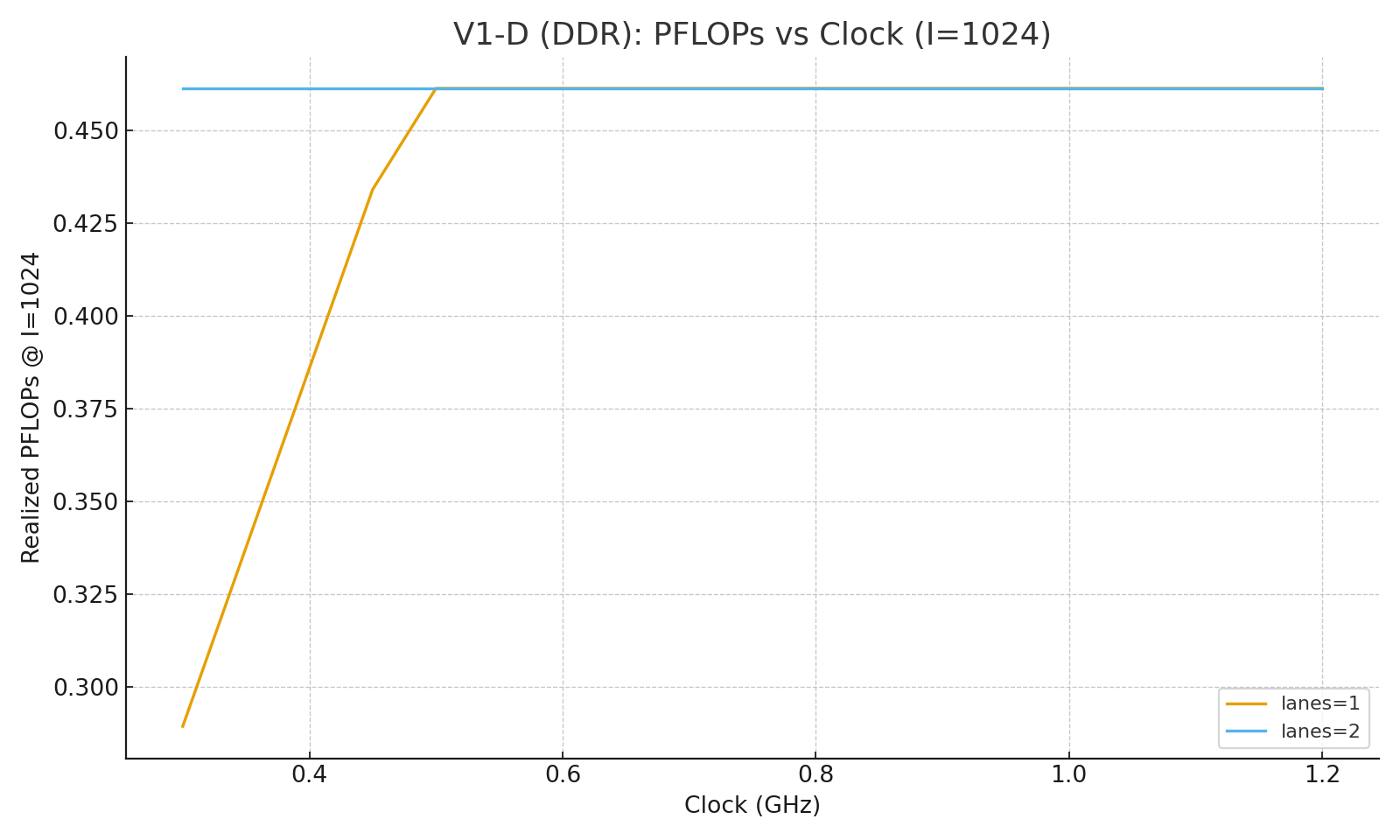
V1‑D (I=256)



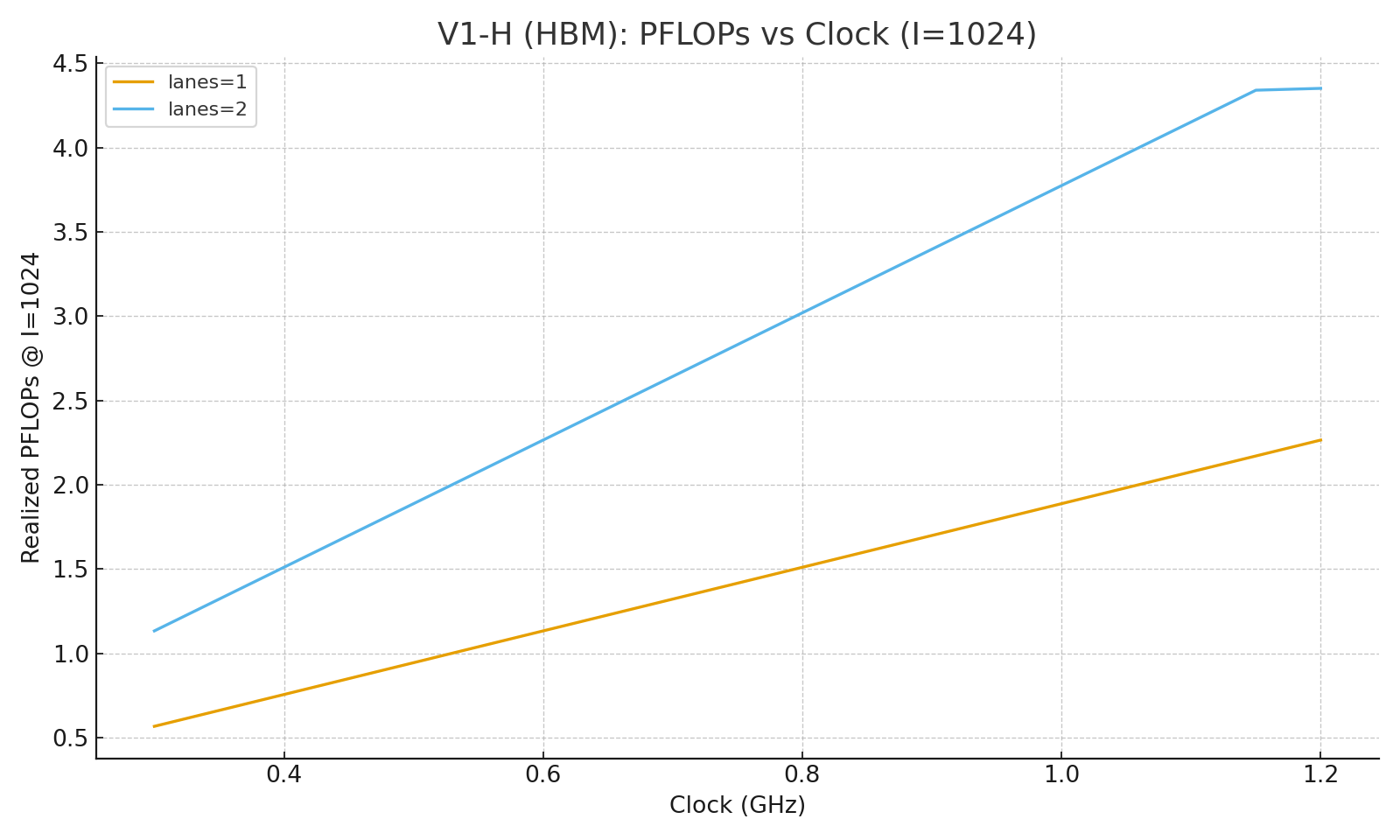
V1‑H (I=256)



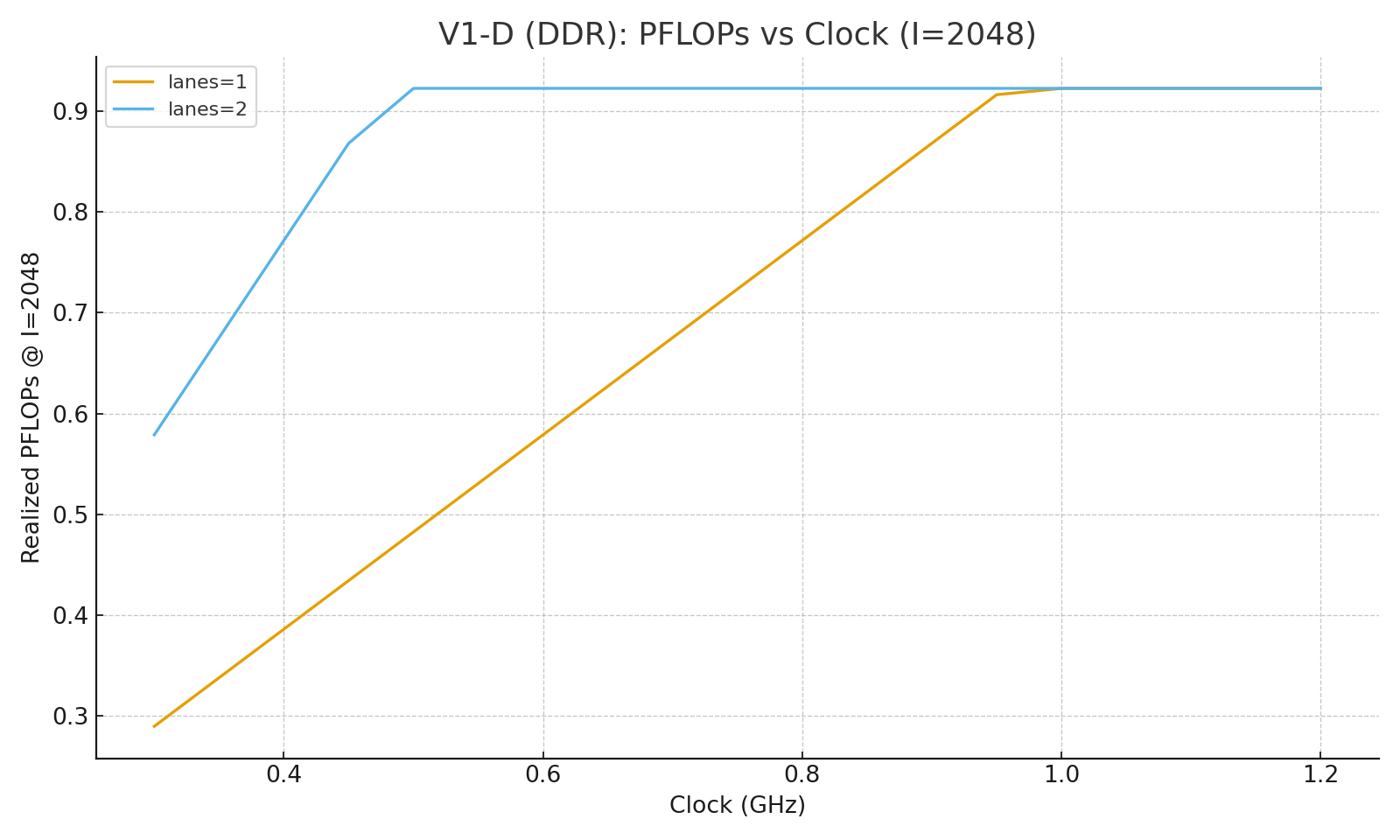
V1‑D (I=1024)



V1‑H (I=1024)



V1‑D (I=2048)



V1‑H (I=2048)

