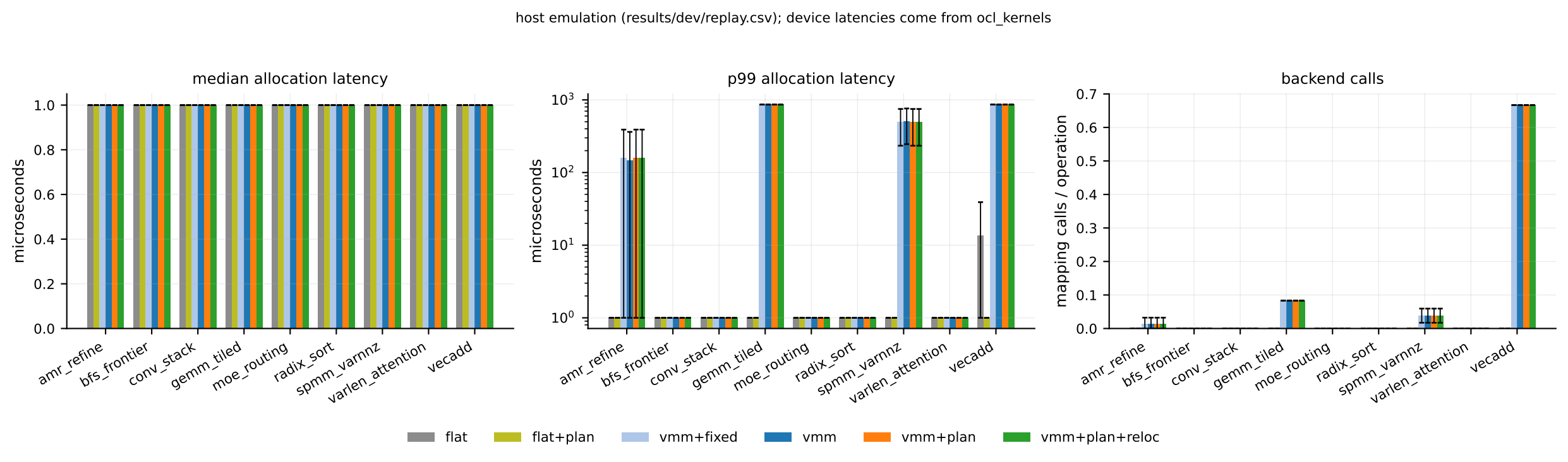
host emulation (results/dev/replay.csv); device latencies come from ocl_kernels
median allocation latency
microseconds
1.0
0.8
0.6
0.4
0.2
0.0
p99 allocation latency
microseconds
10^3
10^2
10^1
10^0
backend calls
mapping calls / operation
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
amr_refine
bfs_frontier
conv_stack
gemm_tiled
moe_routing
radix_sort
spmm_varnnz
varlen_attention
vecadd
flat
flat+plan
vmm+fixed
vmm
vmm+plan
vmm+plan+reloc