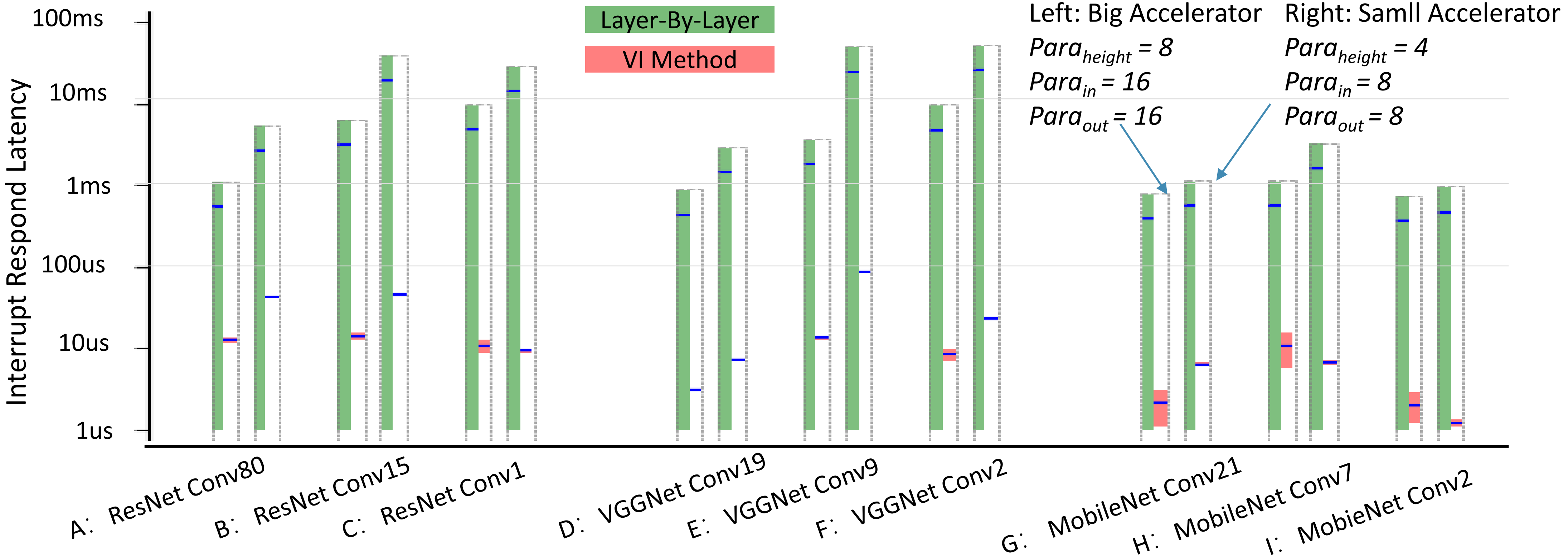
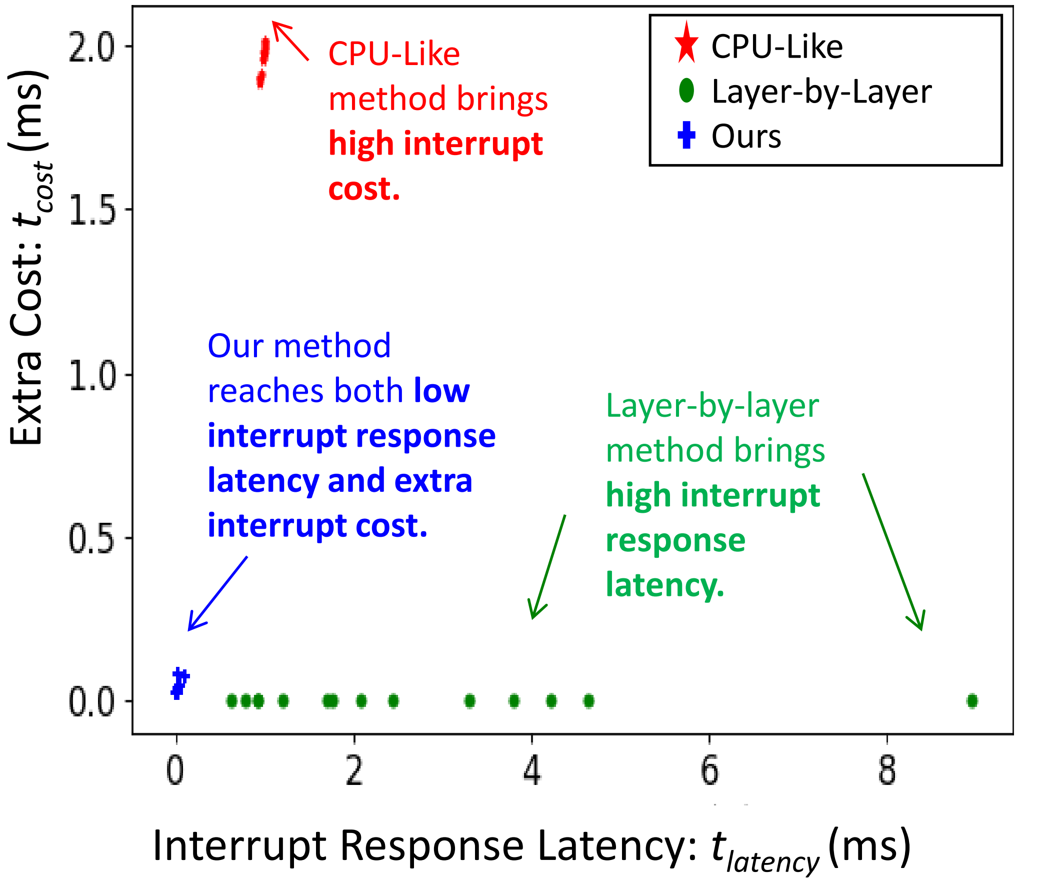


Layer	A	B	C	D	E	F	G	H	I
Network	ResNet101			VGGNet			MobileNet		
W x H	41 x 31	160 x 120	640 x 480	7 x 7	56 x 56	224 x 224	14 x 14	56 x 56	112 x 112
Chin x Chout	1024 x 256	128 x 128	3 x 64	512 x 512	256 x 256	64 x 128	512 x 512	256 x 256	32 x 32
kernel	1	3	7	3	3	3	1	1	3

(a) Latency comparison between layer-by-layer interrupt method and our virtual-instruction (VI) method. We test 9 layers from different networks (Layer A-I), on a big accelerator with larger parallelism (Left bar of each layer) and small accelerator (Right bar of each layer). The green bar shows latency of from best case to worst case of layer-by-layer method. The red bar shows the best to the worst latency of VI method. The blue lines indicates the average latency.



(b) The interrupt response latency ($t_{latency}$) and the extra time cost (t_{cost}) for different interrupt positions in Place Recognition with ResNet101 on the big accelerator with larger hardware parallelism in Fig(a).