key points

sencil is harder than convolution, since it multiple different pattern.

register tiling optimizes memory access by minizing index calculations and leveraging faster register storage. coarening enchangces computational efficiency by enabling threads to handle multiple targets concrrently, thereby increasing parallelism(if per thread is evenly distributed).

1

1.1

1.1.1

the number is 118*118*118

1.1.2

~~that is 10*10*10~~ Effective grid size: 118*118*118, block size:8*8*8. Number of blocks along each dimension:$\lceil \frac{118}{8} \rceil = 15$. So the total number of thread blocks is 15*15*15.

1.1.3

~~the number is 8*8*8~~ Shared memory optimizes memory access but not change the number of thread blocks required for covering the grid. So it is the same as b.:15*15*15

1.1.4

~~the number is 3*32*32~~ $\lceil \frac{118}{32} \rceil = 4$, for covering three layers, the answer is 4*4*3.

1.2

1.2.1

~~32*32*16~~ 32*32*18, z-dimension should take into consider with 2 additional halo layers.

1.2.2

~~32*32*3~~ 32*32*16, in z-dimension 16 consective output planes calculated together, so the processes should be 16 layers.

1.2.3

For memory access, that is 32*32*16*4 bytes. For FLOPs, it is 7(for every point)*16(elements/thread) = 112FLOP.

read operation: 7reads*4bytes; write operation:1write*4bytes.

So totally for (28+4=32)*16=512 bytes.

As a result the answer should be 112/512 ≈ 0.21875FLOPs/Byte.

For allowing significant data reuse,

1.2.4

~~32*32*16*4 bytes~~. for each thread instead of the whold block of its life time, so the number should be 3 layers of pre,curr and next.

1.2.5

32*32*3*4 bytes

The use of register tiling won't effect the need of shared memory.