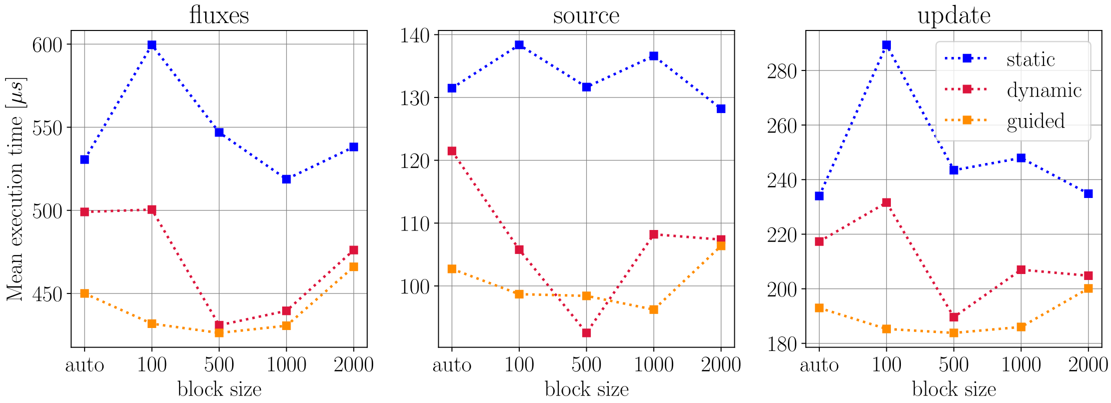

Figure 5.18 – Mean kernel execution time with each scheduling policy

policy suffers significantly from load imbalance and performs much worse than the other two.

5.5 Memory

5.5.1 Roofline analysis

To gain further insights into the bottlenecks in the program and identify what exactly should be optimized, we should first determine whether the program is *compute bound*, meaning the CPU reaches its floating point functional limits, or *memory bound*, meaning the CPU spends most of its cycles waiting on data loads and stores. This is mainly determined by the *data intensity* of the implemented algorithm, which refers to the ratio of floating point operations to the number of bytes moved. The *Roofline model* [77] relates data intensity to the number of floating point operations per second by showing the maximum achievable performance on the system, depending on the peak achievable bandwidth and the peak performance of the processor.

If we can empirically measure the data intensity and the floating point throughput, we can place an algorithm on the Roofline. The roof that intersects with the vertical line drawn from that point allows us to determine whether it is compute bound or memory bound. Furthermore, the vertical gap between the algorithm and the roof helps assess whether there is room for improvement or if the program already pushes the hardware to its functional limits.