Explanation for slide 42: Working set relevant for TLB consideration

The red parts do not show the computation but the working set: parts that are repeatedly accessed and thus we would like to have their page addresses in the TLB.

It is similar to what we did for cache blocking on slide 19 for the mini-MMM: the working set in mini-MMM for the cache in the simple version b) was one block, one row, and one element. Why? The block b is accessed many times in the mini-MMM, one row in block a is accessed several times but when one goes to the second row, the first row is done forever. The element in block c is accessed many times but when one moves to the next one, the first one is done forever. In e) on slide 22 this reasoning is extended to: all of block b + (three) block rows in a + one block in c.

Same reasoning in slide 42, just at a higher level since now we are talking about TLB, so more data can fit. The entire matrix b is accessed many times during the entire computation, namely for each block row of a, so one would hope all the pages fit the TLB. A block row in a is accessed several times (for each block column of b), but once one moves to the second block row of a, the first is never accessed again etc.