5. Explain how the code becomes faster due to cache hits.

When some address is accessed, a large chunk of memory (close to each other) is fetched (not just that particular memory location) and is cached. So, when we do a block-by-block transpose, the process becomes faster as the data will fetch closer to each other in memory in one go (cache hit). However, in doing a brute force transpose, there will be cache miss that will result in a slower process.