3.f) In a couple sentences, explain the difference that you see in the times for mmul3 and mmul4 when running on an Euler compute node. Make sure to discuss the hardware design and access patterns that cause this difference.

Ans) According, mmul3 is roughly 3 times faster than mmul4. This is caused because in mmul3 when data is accessed, entire Cache Block is fetched into the Cache. Since, the Matrix Data of both A and B are in row order, hence when you request for one data, due to spatial locality within Cache block, the neighboring pixel is stored in the Cache too. Hence, the access for data within a row is faster since most of the row data is already present in Caches.

The performance is slower in case of mmul4 because both matrices are stored in Column major order. Here, for each individual pixel, you need to repeated fetch entire Cache block. This can also leads to higher cache evictions which leads to repeated Data accesses from the upper level storage elements(L2 Cache and so on).