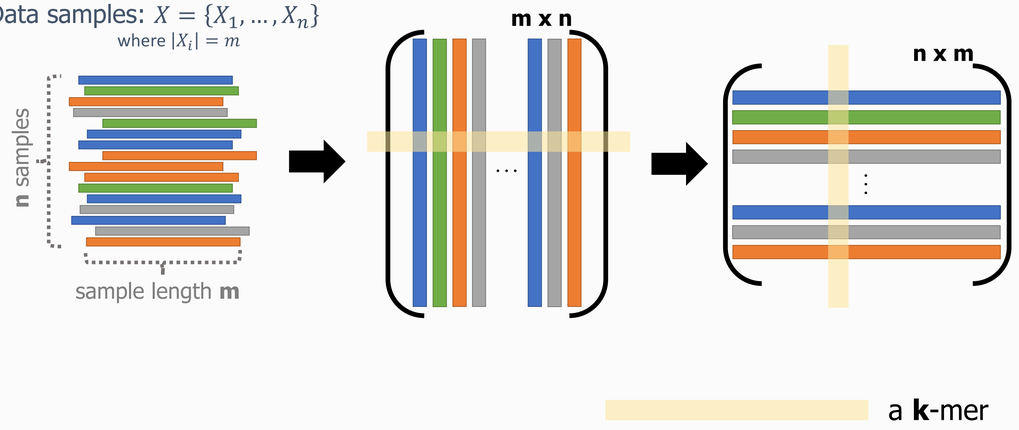
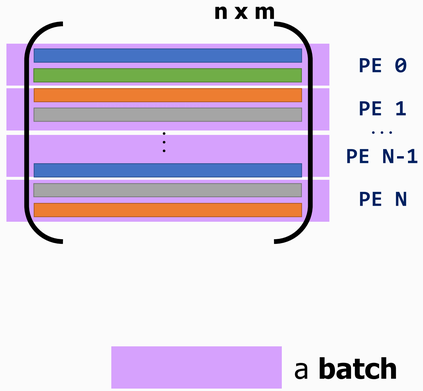
Distributed, Scalable, and Asynchronous Actors Approach to Jaccard Similarity for Genome Comparisons

1 create indicator matrix from data samples and transpose matrix

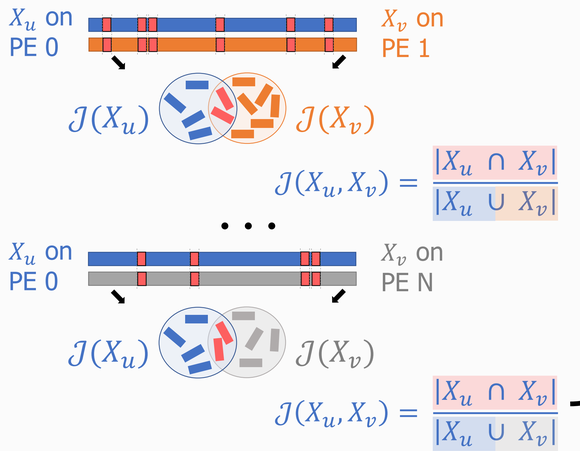
Data samples: $X = \{X_1, \ldots, X_n\}$ where $|X_i| = m$

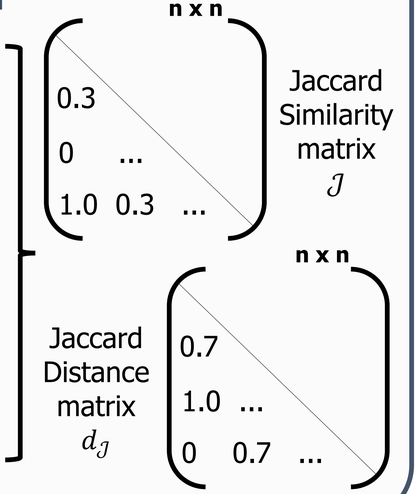


2 create batch for each PE



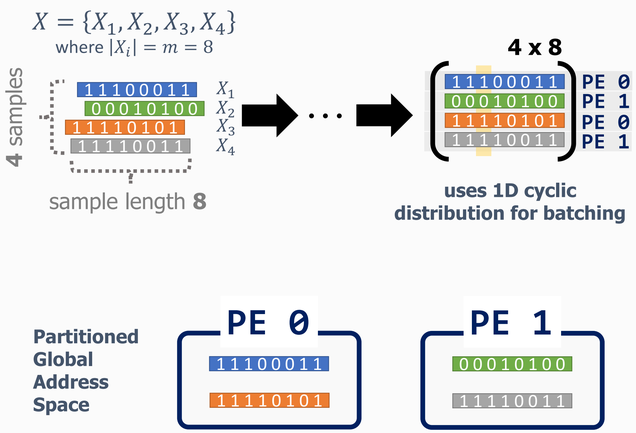
3 async. communication and computation to calculate Jaccard similarity



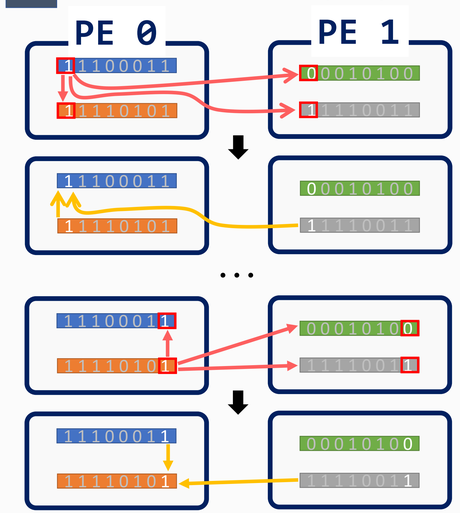


Example Execution Flow using ($m = 8$, $n = 4$, $\#PEs = 2$)

1, 2

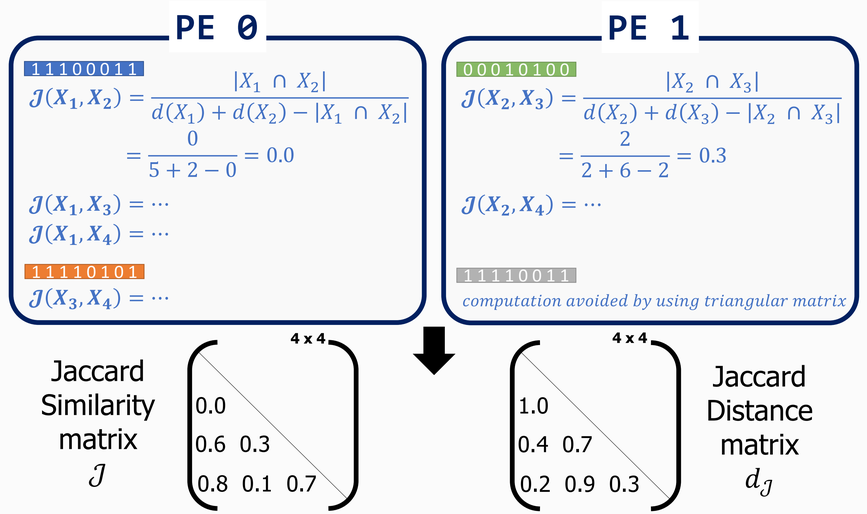


3 execution from the perspective of **PE0** (assume SPMD, all PEs execute same steps)



INTERSECTION CALCULATION DONE, INVOKE BARRIER BEFORE EXECUTING NEXT PHASE

EXTENDED BARRIER



DESCRIPTIONS:

- a **k**-mer
- Send an asynchronous message to other PEs to check if the same k-mer is present in any of the other data samples [CODE: **Listing 1, L32**]
- Highlights the k-mer of interest in the comparison for each data sample [CODE: **Listing 1, L34**]
- Send an asynchronous message back to sender PE to update a local intersection counter since a common k-mer was found [CODE: **Listing 1, L36**]
- Extended barrier to wait for completion of sending and receiving async. messages before proceeding [CODE: **Listing 1, L44**]