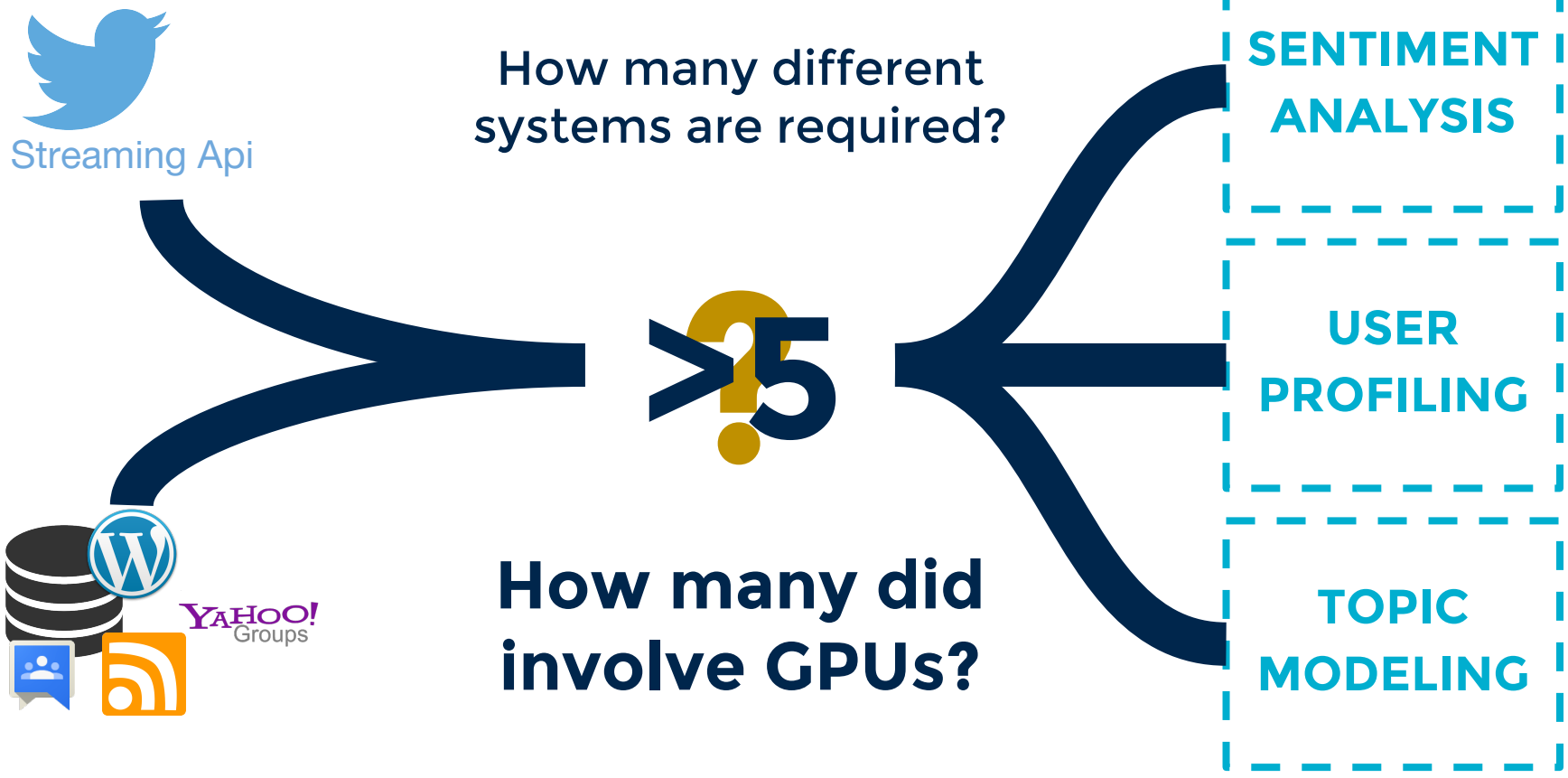
EURA NOVA

GPU TECHNOLOGY CONFERENCE

Maximize the Performance of your Cluster: Marrying GPUs and Dataflow Graph Processing

GPU TECHNOLOGY CONFERENCE 2015
SAN JOSE, CA, USA
MARCH 2015

TRAN Nam-Luc, Engineer @EURA NOVA Research & Development


Streaming Api
How many different systems are required?
>5
SENTIMENT ANALYSIS
USER PROFILING
TOPIC MODELING
YAHOO! Groups
How many did involve GPUs?

THE BIG PICTURE

We had computing clusters.

Now, we have clusters and GPUs.

How to maximise the overall usage and performance?

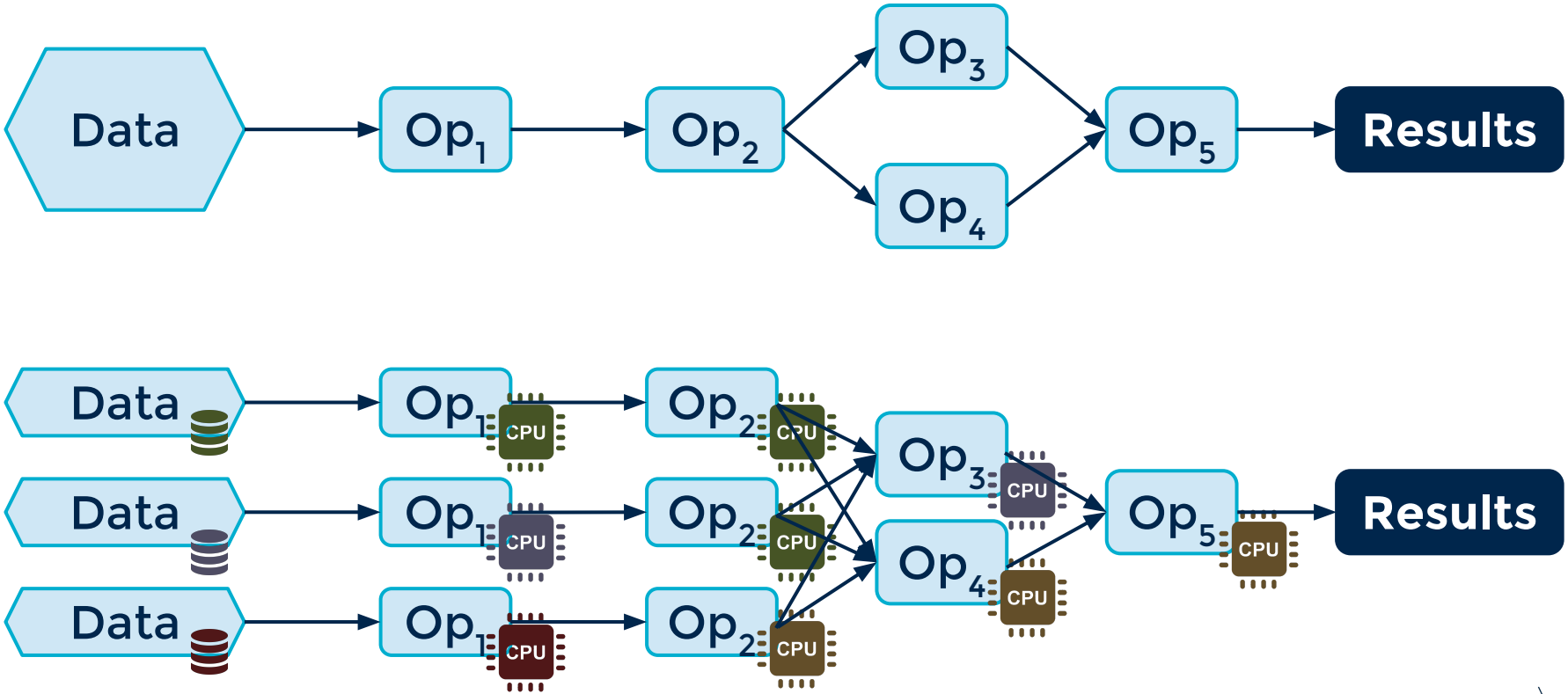
OUTLINE AND CONTRIBUTIONS

1. **A REAL WORLD CASE**

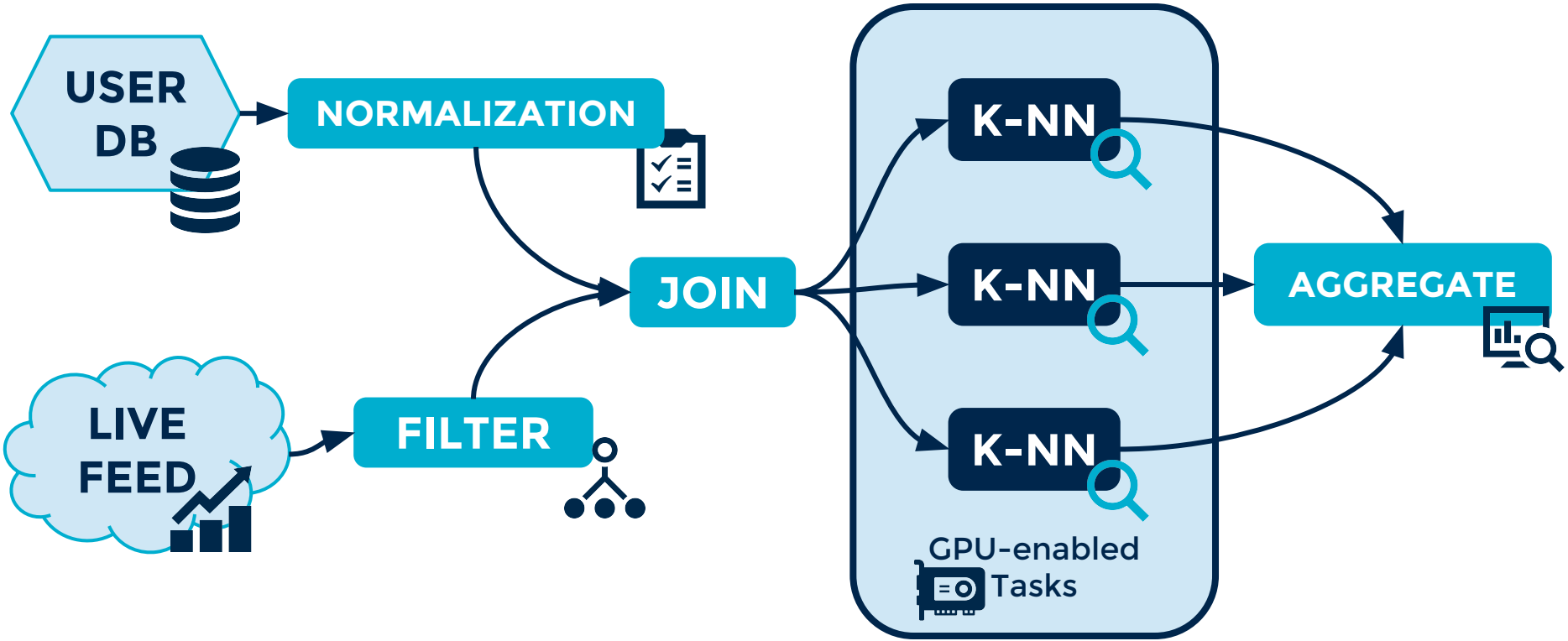
2. **THE DFG PROCESSING MODEL**
   Support for GPU tasks

3. **THE HETEROGENEOUS SCHEDULING MODEL**
   Generalization for GPU resources

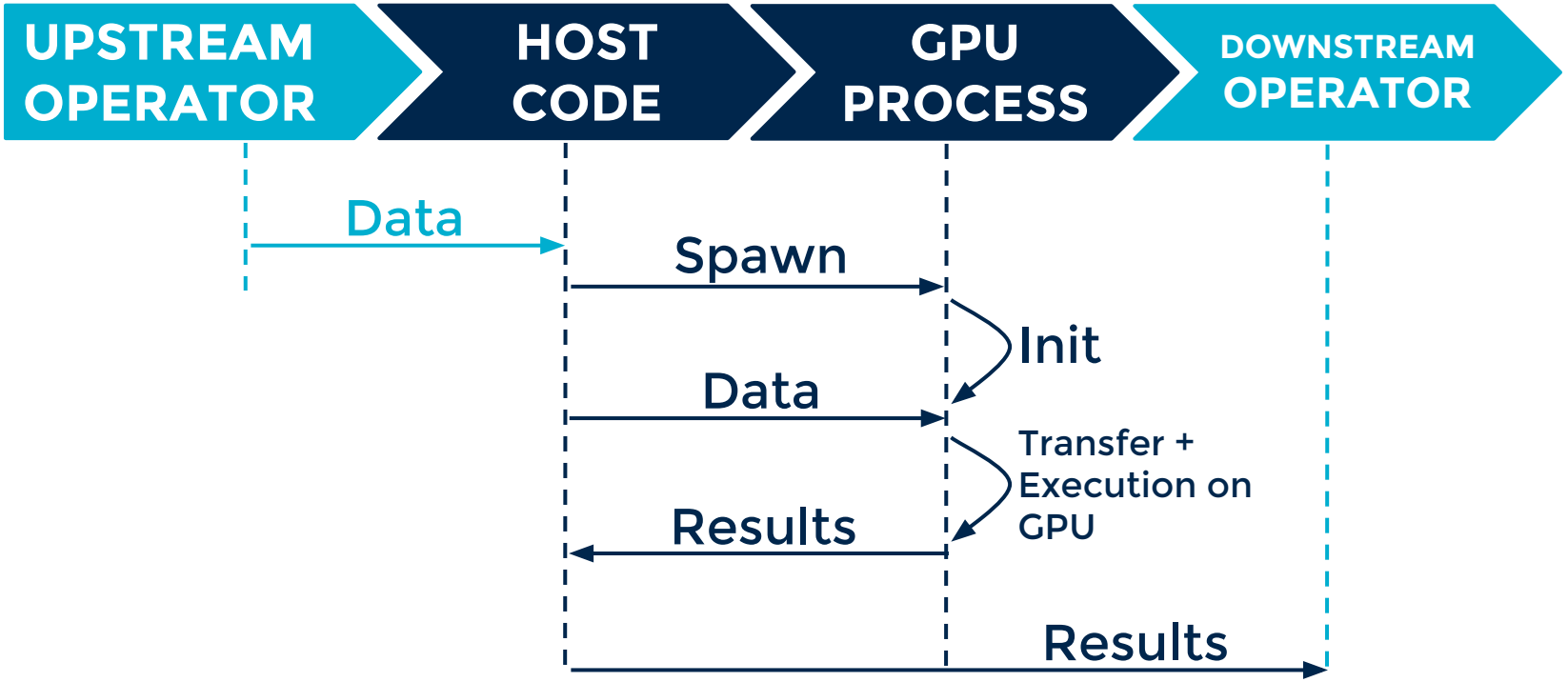
DFG PROCESSING FRAMEWORKS



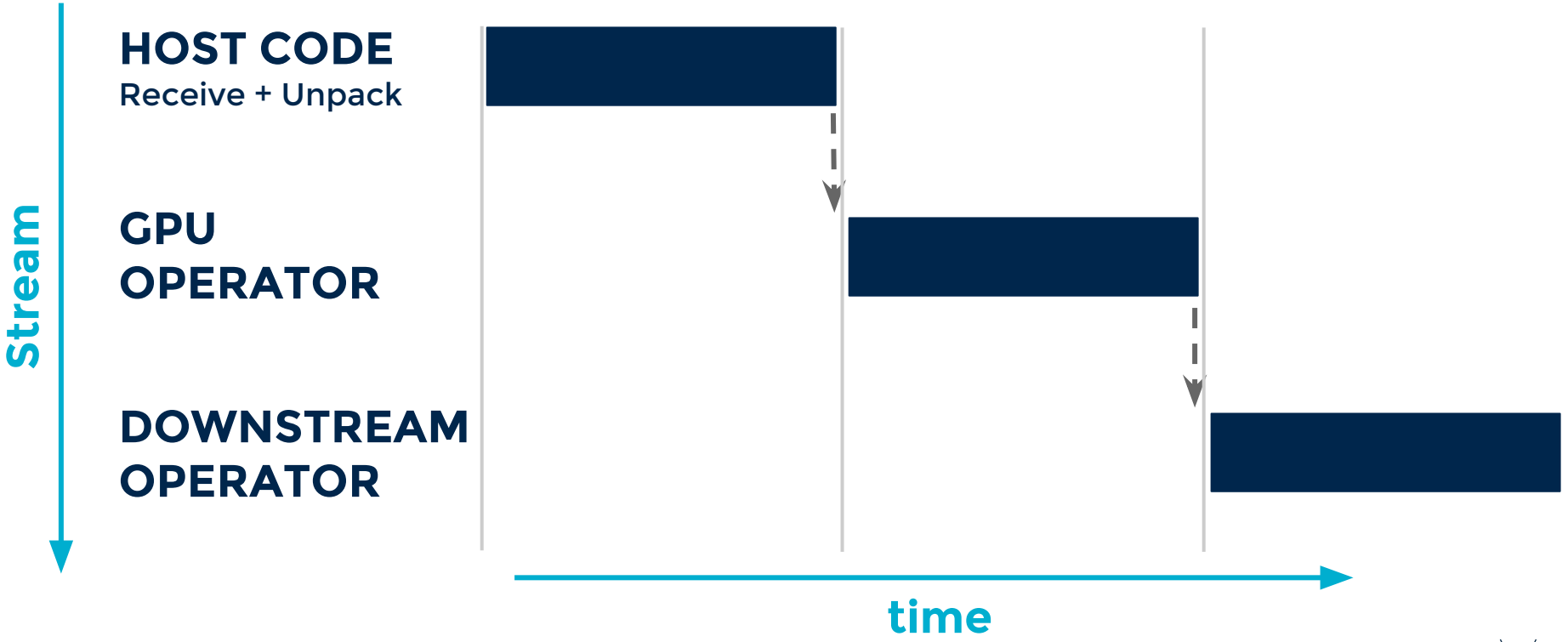
GPU TASKS IN DFG JOBS



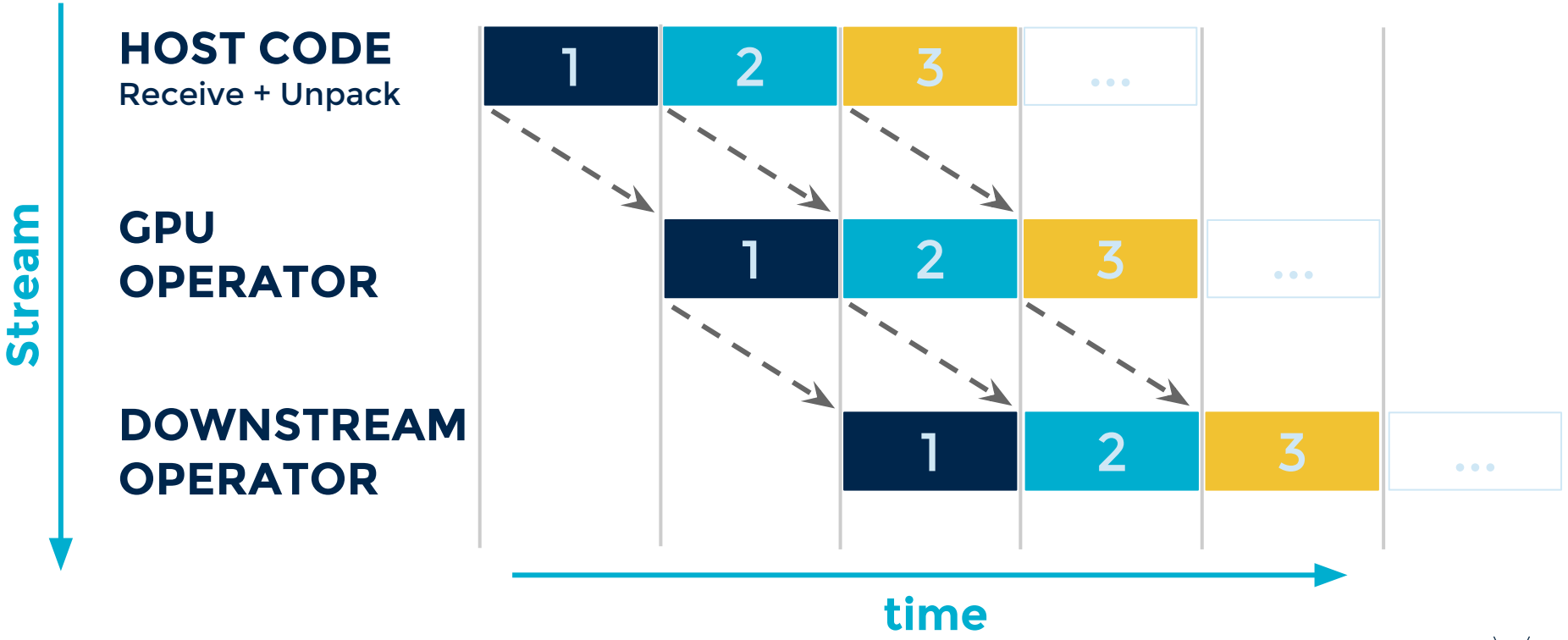
GPU OPERATOR INTERNALS



BATCH OPERATOR



STREAMING OPERATOR



BATCH vs STREAMING

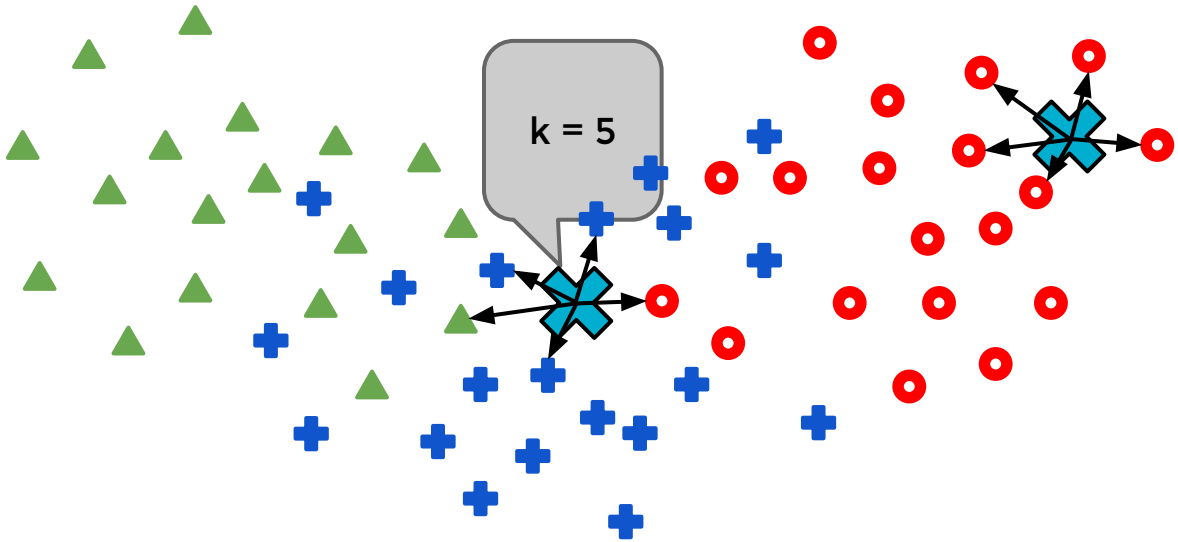
BATCH

- ➕ Massive parallelism
- ➖ Memory bound
- ➖ Coarse grain overlap

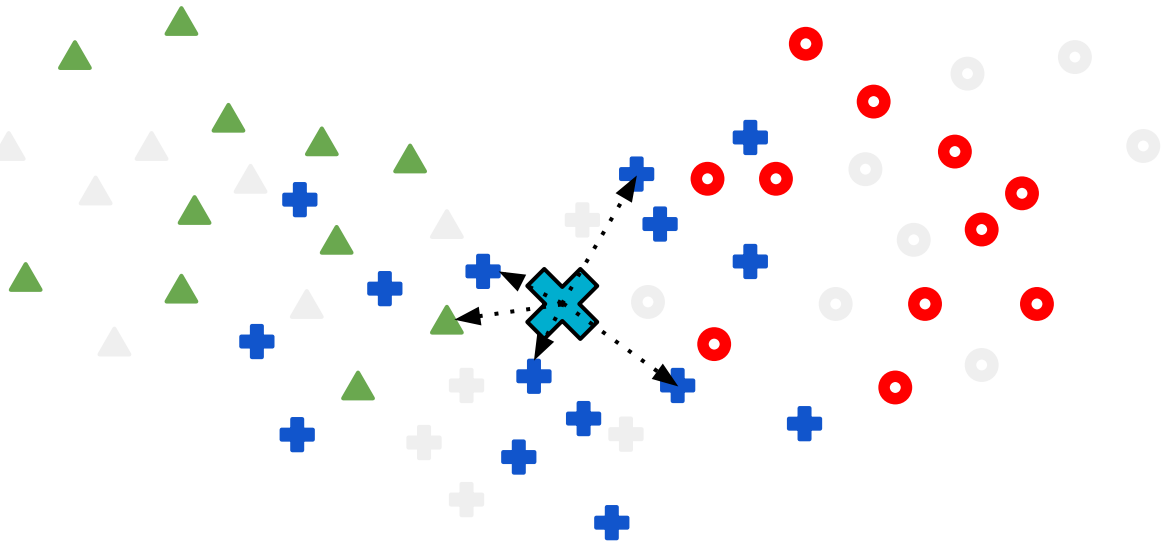
STREAMING

- ➕ Fine grain overlap
- ➕ Faster results
- ➖ Less massive parallelism

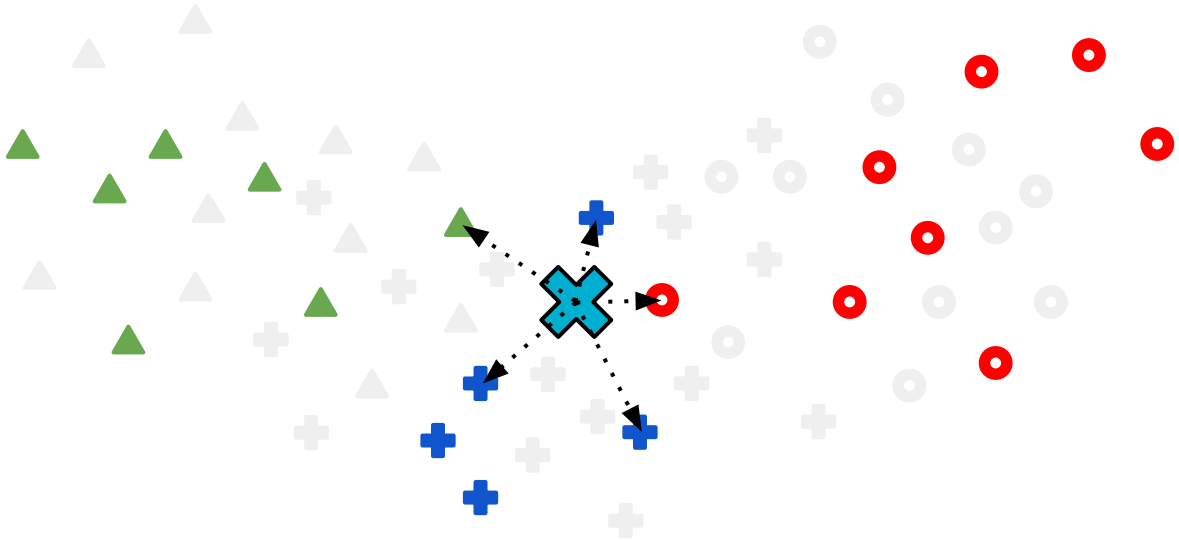
ILLUSTRATIVE JOB: K-NN



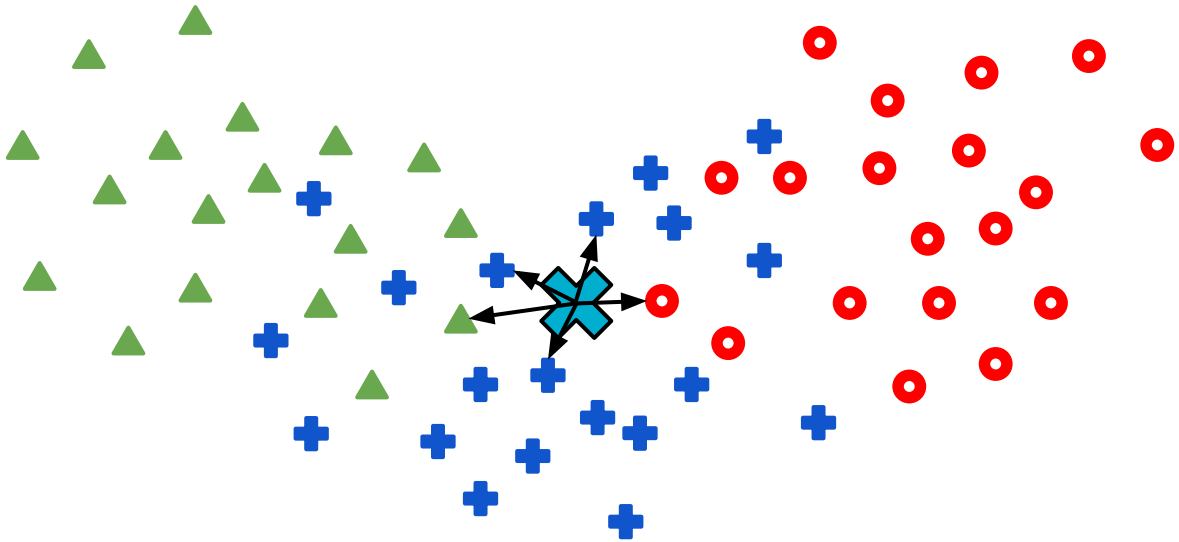
K-NN DISTRIBUTION STRATEGY



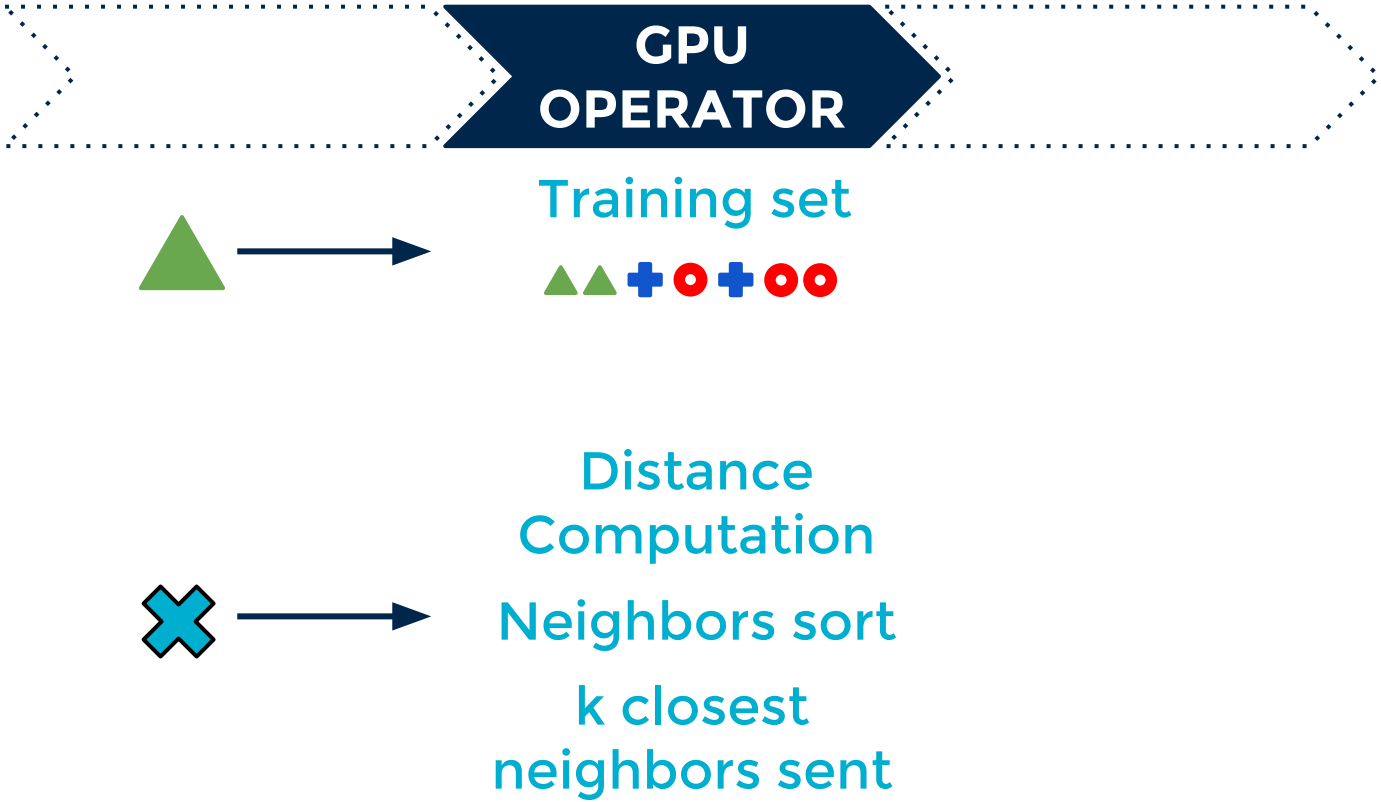
K-NN DISTRIBUTION STRATEGY



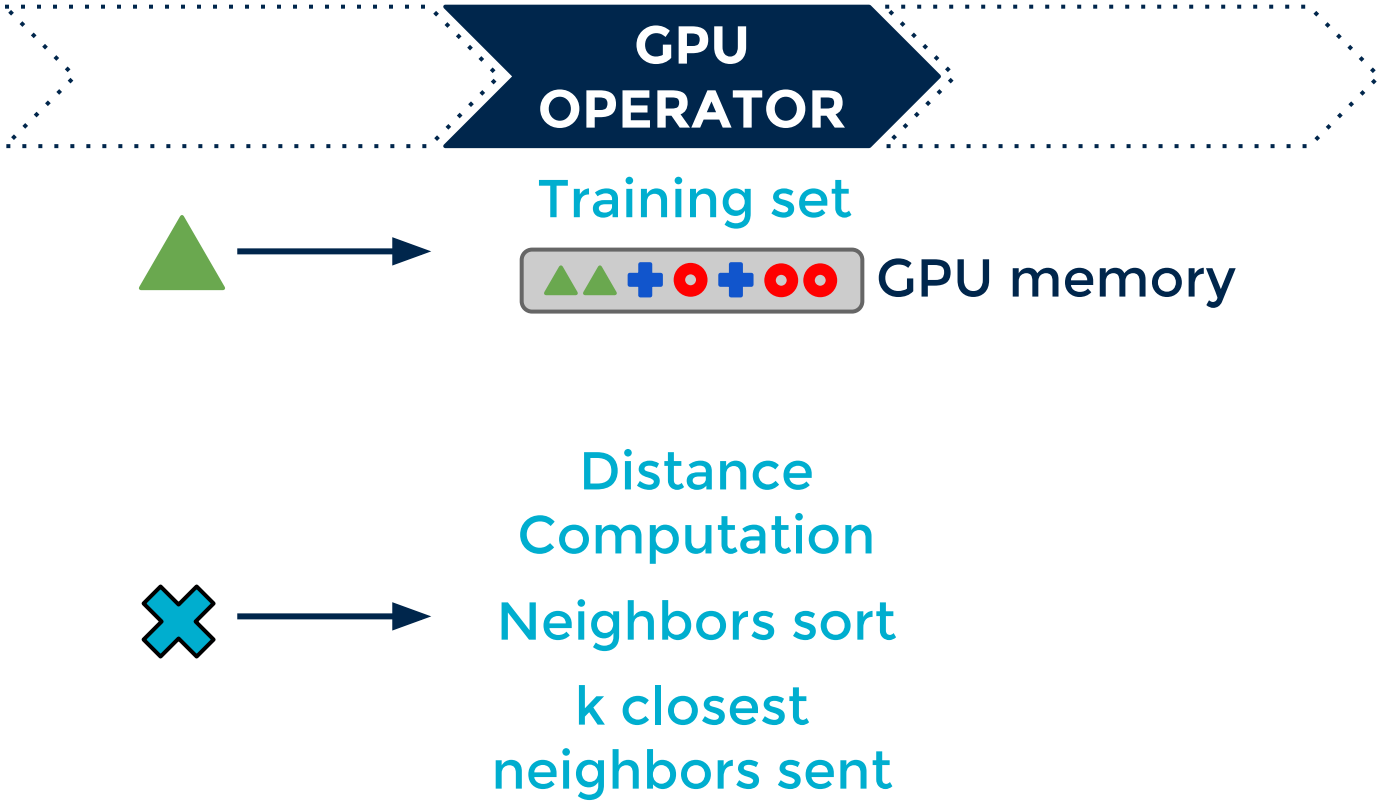
ILLUSTRATIVE JOB: K-NN



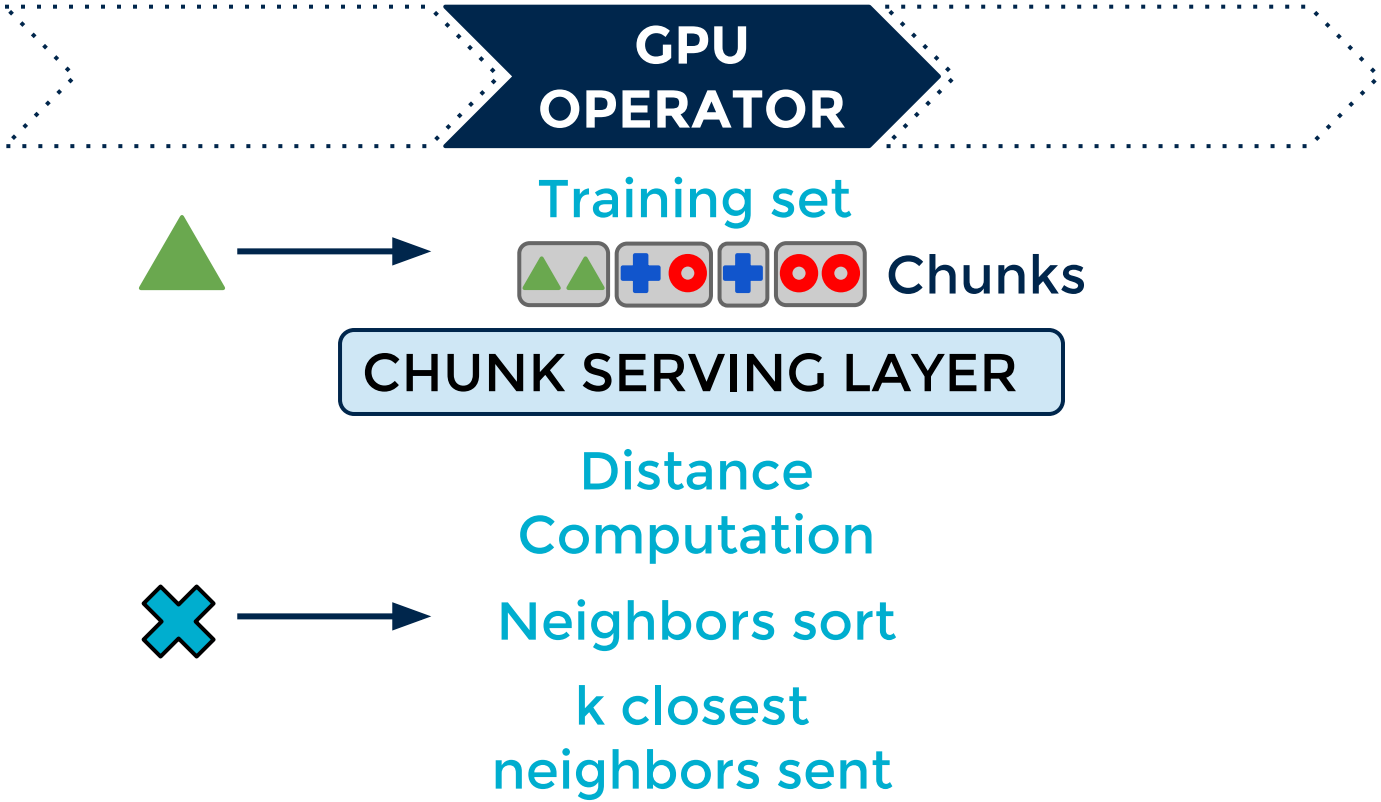
K-NN GPU OPERATOR



K-NN GPU OPERATOR



K-NN GPU OPERATOR



BATCH vs STREAMING

Tradeoff:
MINI-BATCH

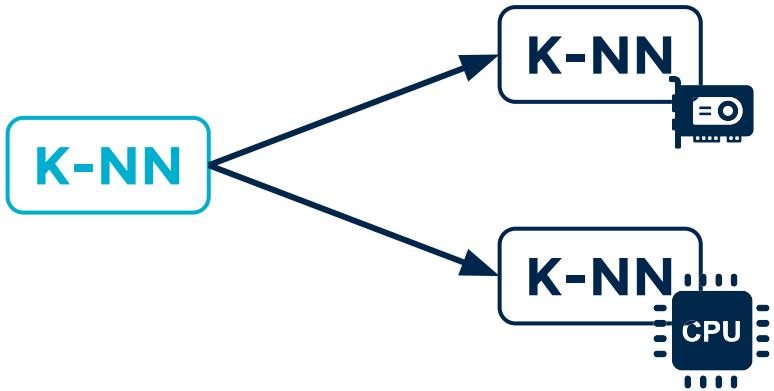
Thread block size: 8
Chunk size: 1200 elements, 23 features
Stream size: 2

Hardware: i5 3.4 Ghz 16Gb RAM +
Tesla K20

HETEROGENEOUS SCHEDULING

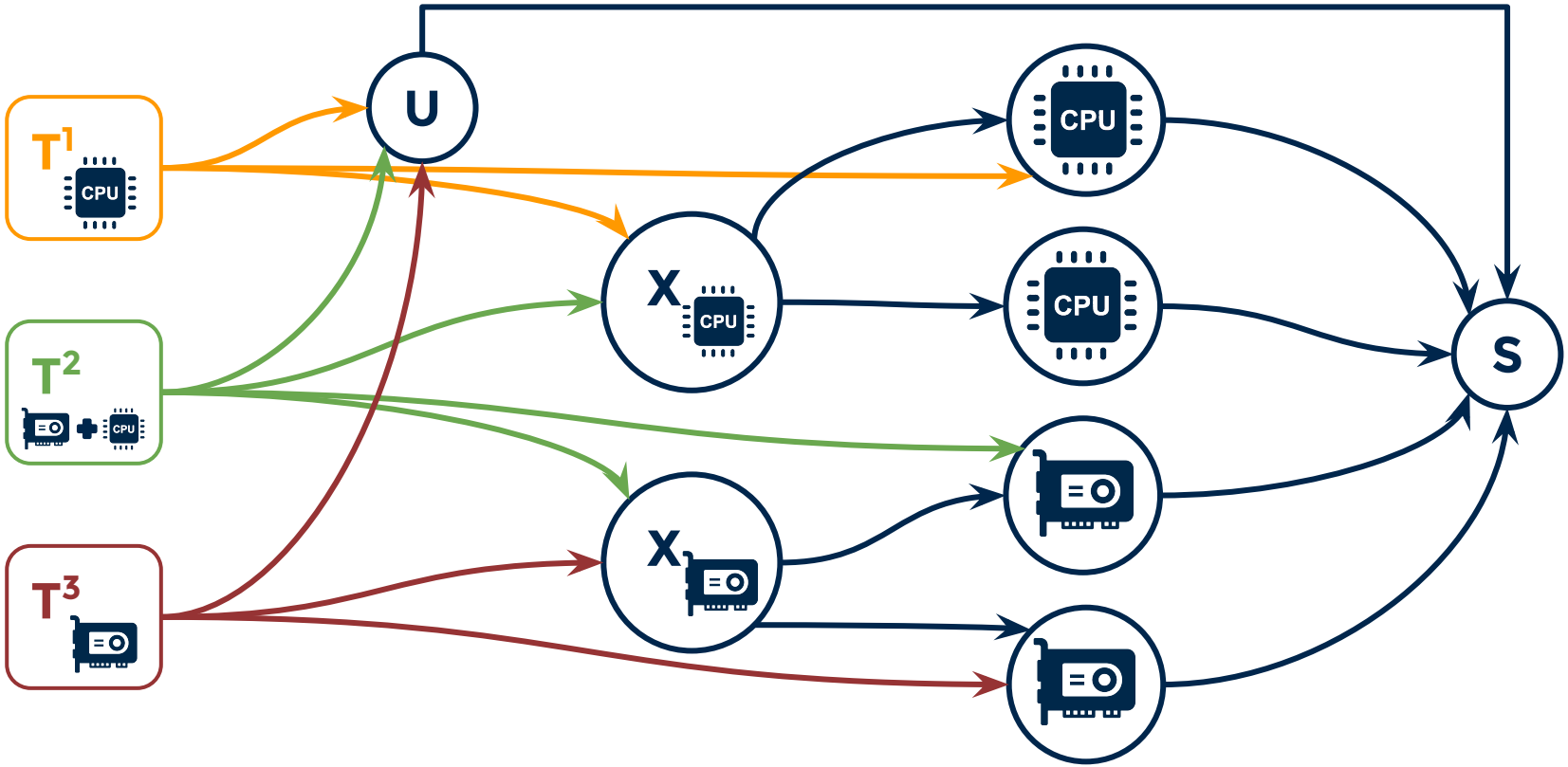
We have CPU and GPU resources in the cluster.

Operators can have preference over one type of resource.

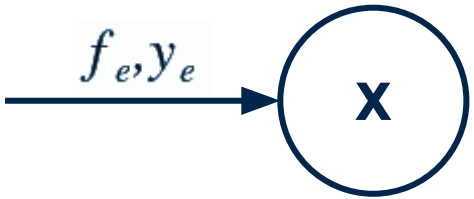


“x times faster than the CPU version!”

THE SCHEDULING MODEL



THE SCHEDULING MODEL



Flow and supply $f_e \leq y_e$

Capacity $\sum_v \varepsilon_v = 0$

Feasible flow $\varepsilon_v + \sum_{e \in I_v} f_e = \sum_{e \in O_v} f_e$

Scheduling solution $\min \sum_e f_e \cdot p_e$

CONCLUSION

We had computing clusters.

Now, we marry clusters *and GPUs.*

EURA NOVA

GPU TECHNOLOGY CONFERENCE

THANK YOU!

@Namux #GTC15

Please complete the Presenter Evaluation sent to you by email or through the GTC Mobile App. Your feedback is important!