COSTREAM: Learned Cost Models for Operator Placement in Edge-Cloud Environments

Roman Heinrich, Carsten Binnig,
Harald Kornmayer, Manisha Luthra

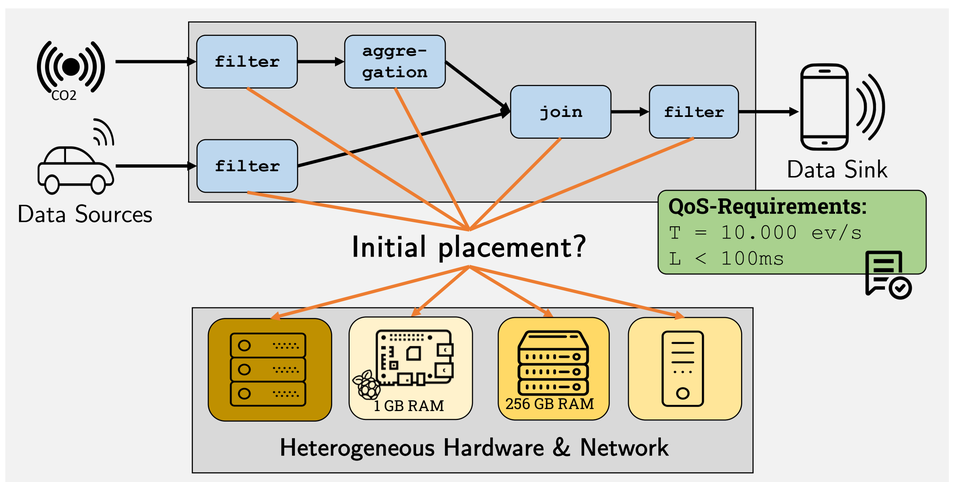
ICDE 2024

TECHNISCHE UNIVERSITÄT DARMSTADT

DHBW Duale Hochschule Baden-Württemberg Mannheim

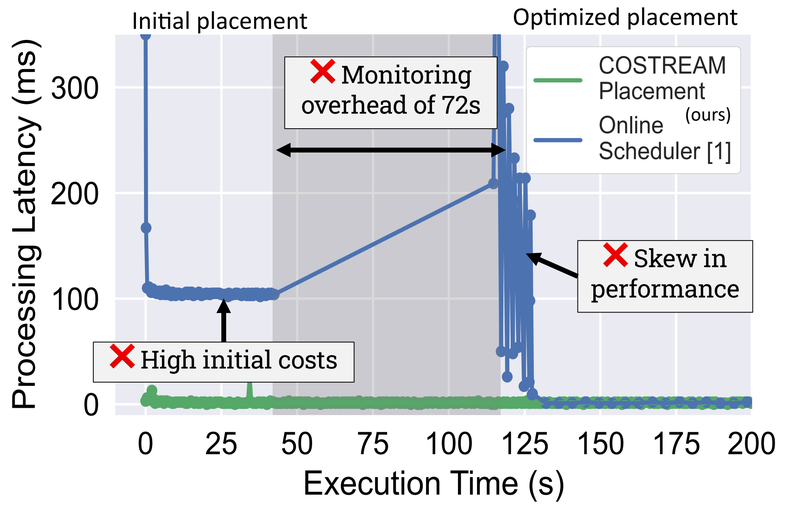
dfki ai

Stream Processing in IoT-Scenarios & Initial Operator Placement



Motivation

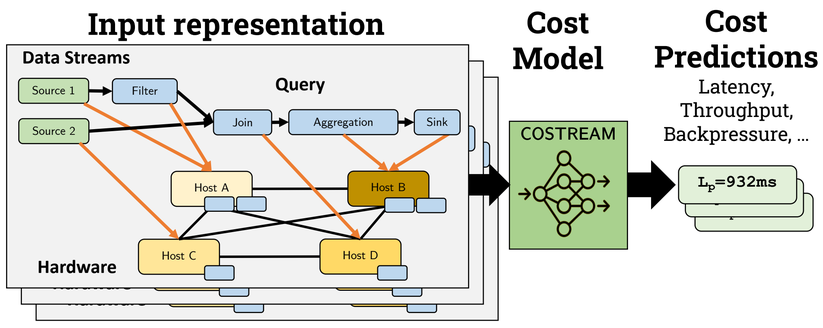
- Stream Processing queries are often executed in a IoT or Cloud-Edge scenario where the resources are **heterogeneous** in terms of compute and network
- A crucial challenge is to find an **initial placement** of streaming operators to these resources.
- Suboptimal placements lead **to high initial** costs like latency and throughput and **require rescheduling downtimes**



Main Contributions & Core Ideas

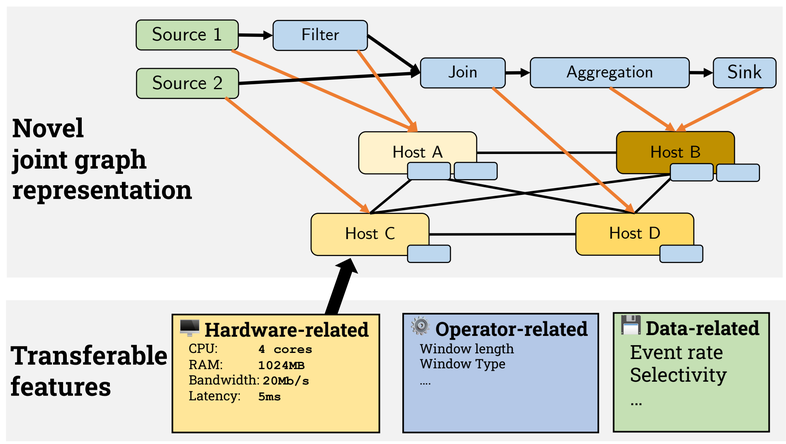
Overview

COSTREAM is a **learned cost model** that predicts the execution costs for an initial operator placement on heterogeneous hardware.
COSTREAM paves the road **for cost based optimization** of stream processing systems. In this work, we use cost estimates to reason about the optimized initial operator placement of a streaming query.



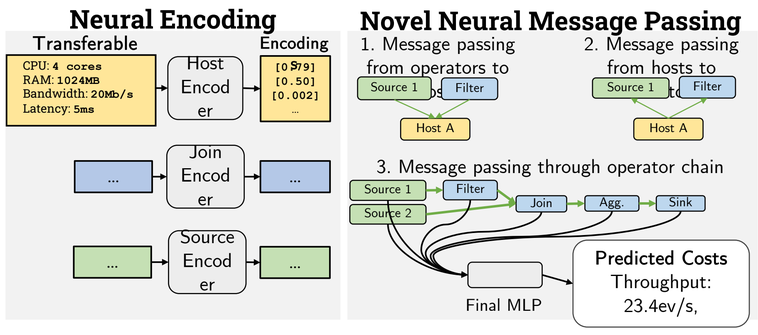
Transferable Input Representation

The query operators, hardware resources and data streams are modeled in a **joint graph** and assigned with **transferable features.**



Learning Placement Costs with Graph Neural Networks

Placement costs are learned with a novel two-stage learning approach that first embeds the node features into hidden states and then applies a **novel neural message passing scheme.**

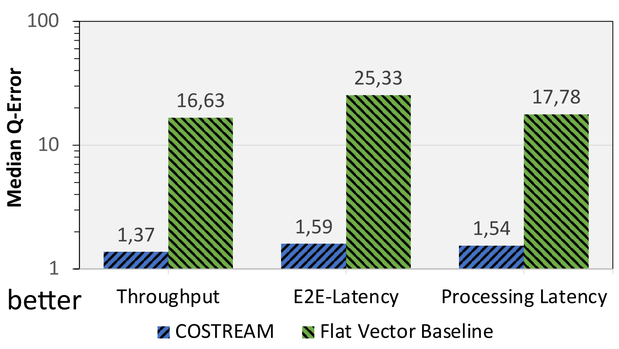


Experimental Evaluation

How good are cost predictions from COSTREAM?

- **Method:**
  Test predictions for **unseen hardware** that differs from initial training data.
- **Example**
  Training range - RAM:
  `1, 2, 4, 8, 16, 24, 32`
  Evaluation range - RAM:
  `1.5, 3, 6, 12, 20, 28`
- **Metric**:
  Deviation of real and predicted costs with median Q-Error:

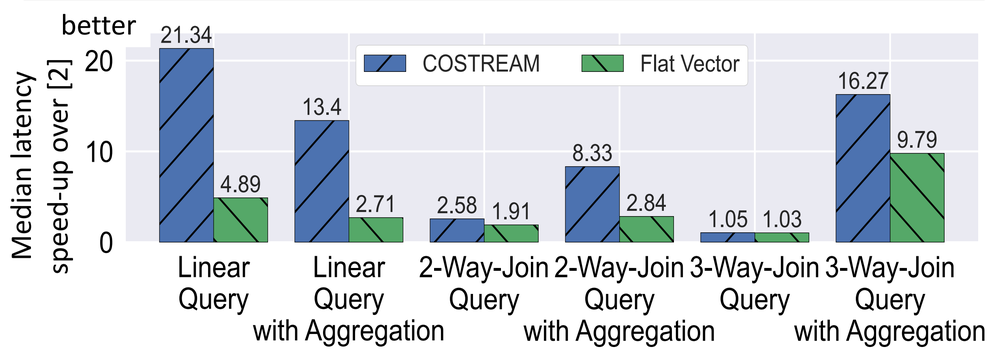
$$Q(x, x') = \max\left(\frac{x}{x'}, \frac{x'}{x}\right)$$



✅ Accurate predictions for unseen hardware

How good are the initial placements provided by COSTREAM?

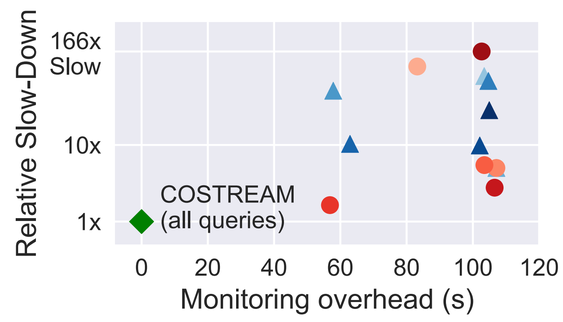
Method: Optimize initial heuristic placements based on [2] with **COSTREAM** and **Baseline**



✅ COSTREAM returns placements with high speed-ups across query types

What are benefits of cost-based placement optimization?

Method: Compare initial placement from COSTREAM versus an online scheduling approach [1]



Event Rates (ev/s): 100, 200, 400, 800, 1600, 3200, 6400
Selectivities: 0.1, 0.25, 0.5, 0.75, 0.9, 1.0

✅ High initial speed-ups of up to 166x

✅ Avoiding monitoring overhead of up to 120s

[1]: Aniello, *et al.* "Adaptive online scheduling in storm." *DEBS* 13.
[2]: Chaudhary et. al „GOVERNOR: „, *EDBT'20*

Paper