# Graph Neural Network-based Tracking as-a-service

Xiangyang Ju[1], Elham E Khoda[2], Andrew Naylor[3], Haoran Zhao[2], co-authored with Paolo Calafiura[1], Steven Farrell[3], Shih-Chieh Hsu[2], William McCormack[4], Philip Harris[4], Dylan Rankin[5], Yongbin Feng[6]

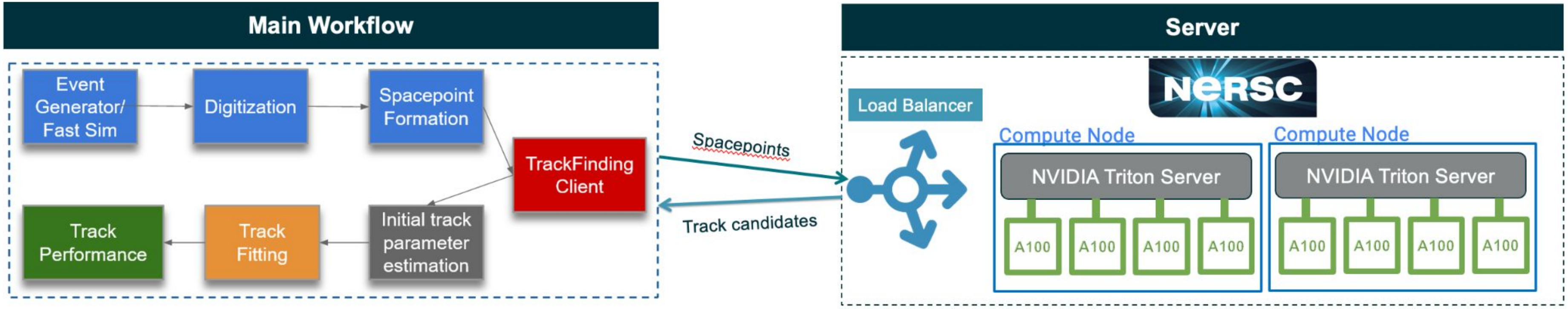
1. LBNL
2. University of Washington, A3D3
3. NERSC
4. MIT
5. University of Pennsylvania
6. FNAL

# GNN-based Tracking as a service

Why tracking as a service?

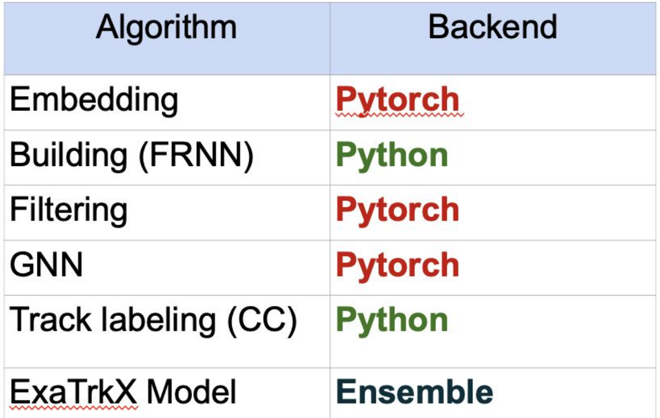
1. Factorize out ML framework
   - ○ Easy support for different ML frameworks & models
2. Factorize out algorithm scheduling
   - ○ ML models can be deployed on different coprocessors simultaneously and easily
3. Portable solution to supporting different coprocessors. No need for client to rewrite code for specific languages

4. Allow access to remote AI accelerators

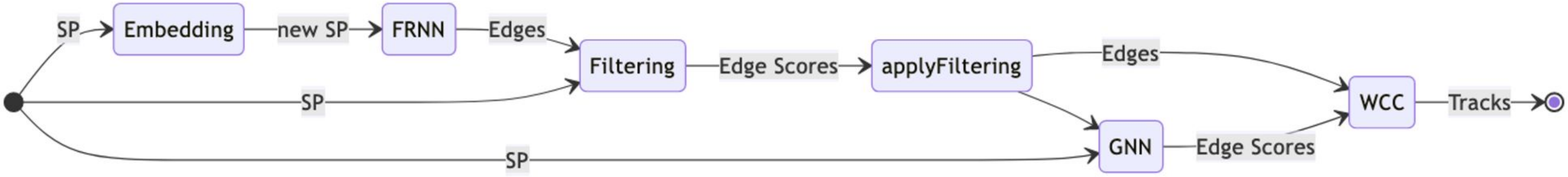


# Ensemble Backend

- GNN-Based Tracking is a complex workflow, consisting of 5 discrete sub-algorithms
- Ensemble scheduling uses greedy algorithms to schedule each algorithms (see the flow chart)
  - **Pros**: directly use existing Triton inference backends
  - **Cons**: little control with the data flow and algorithm scheduling, increasing the IO operations and latency



| Algorithm | Backend |
|---|---|
| Embedding | **Pytorch** |
| Building (FRNN) | **Python** |
| Filtering | **Pytorch** |
| GNN | **Pytorch** |
| Track labeling (CC) | **Python** |
| ExaTrkX Model | **Ensemble** |

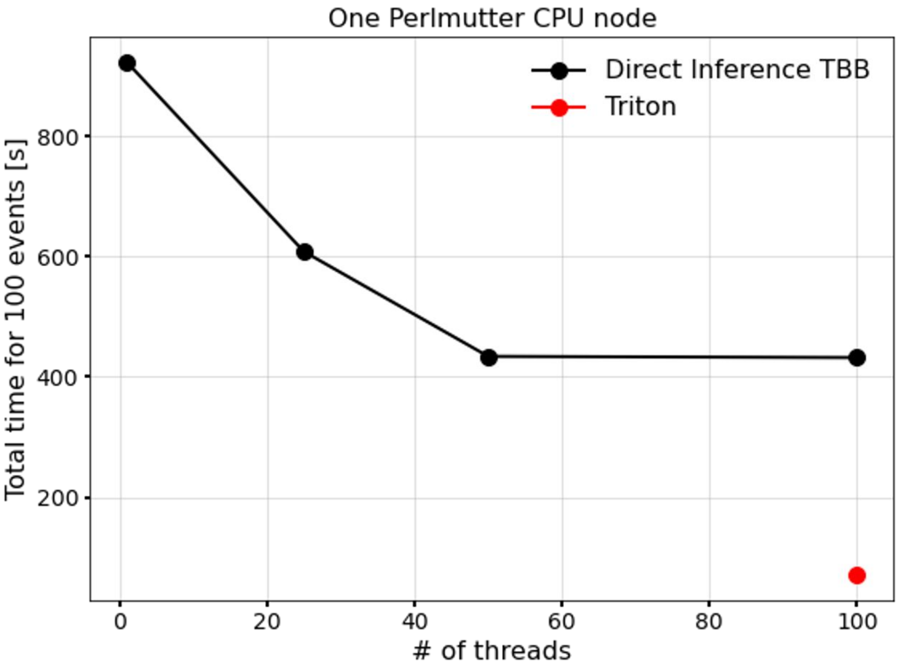


# Customized Backend

Customized backend provides means to receive requests from and send outputs to the client. ***Pros*** : low overhead, full control of data flow and devices; ***Cons*** : need to write user’s own inference code

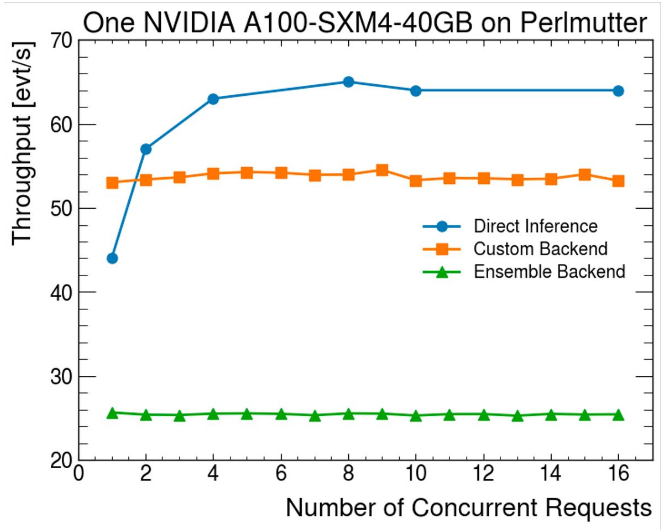
We build customized backends for the CPU-only and the GPU-only ExaTrkX inference service.

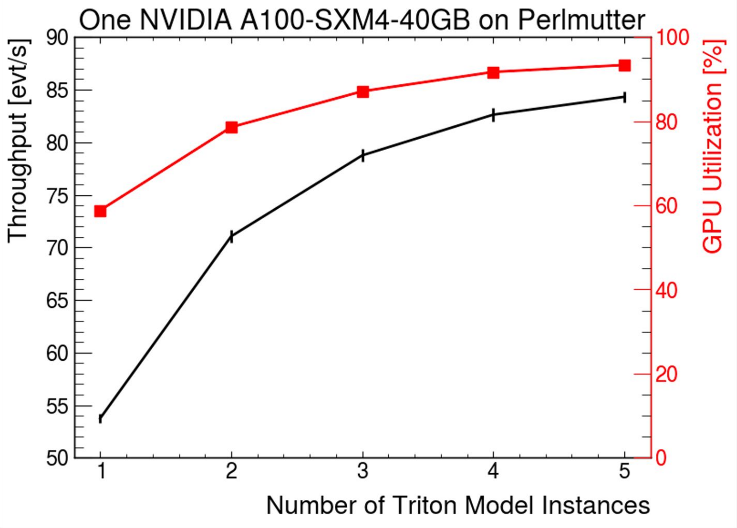
# CPU-based GNN Tracking Service



Triton Server knows how to better utilize CPU resources than a simple TBB scheduling

# GPU-based GNN Tracking Service





- Increasing Triton model instances increases the GPU utilization and throughput
- Customized backend is better than Ensemble model for complex workflow like the GNN-based Tracking
- Direct inferences require higher concurrency to reach maximum throughput