

境外研修答辩

致远学院 2015级 ACM班 孔冰玉

Research Experience

- Nexus-A GPU Cluster for Accelerating Neural Networks for Video Analysis
- This project is to address the problem of serving Deep Neural Networks (DNNs) efficiently from a cluster of GPUs.



(a) Traffic monitoring



(b) Game stream discovery

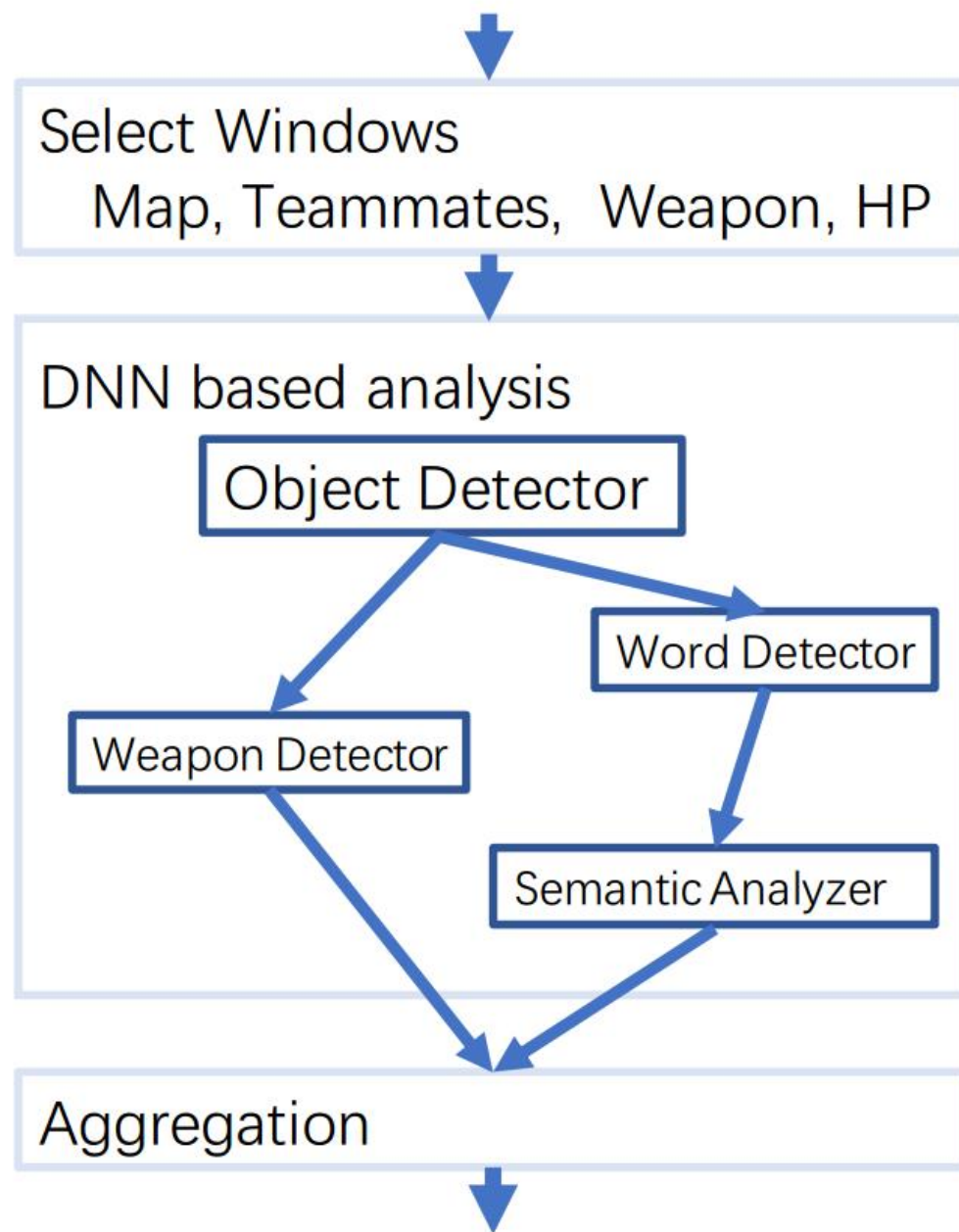
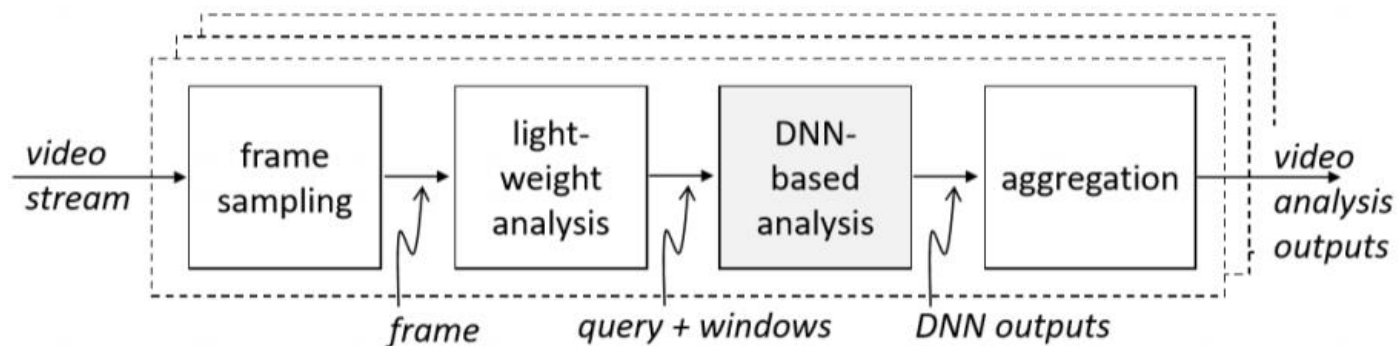


(c) News indexing

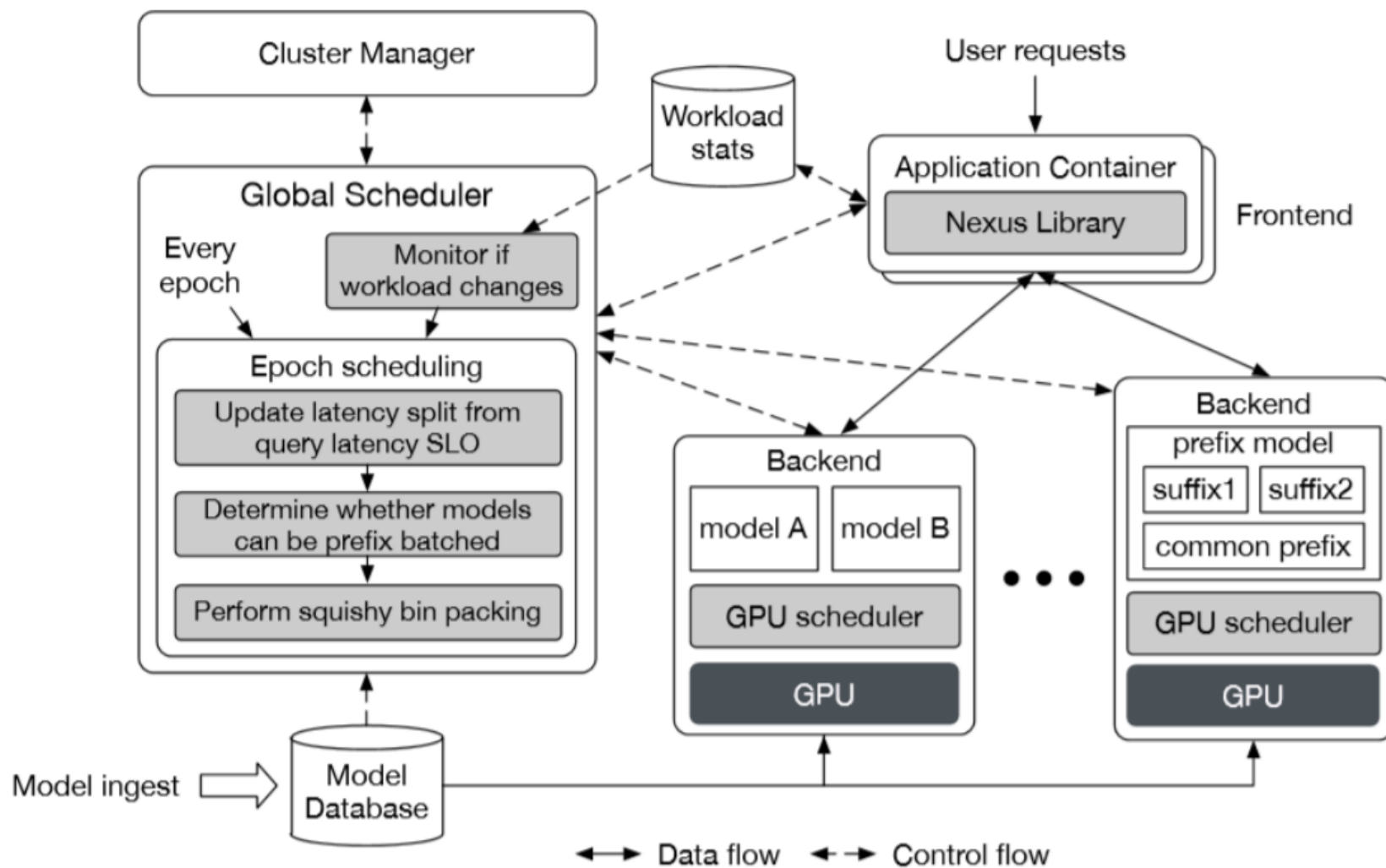


(d) Building access

Question Definition

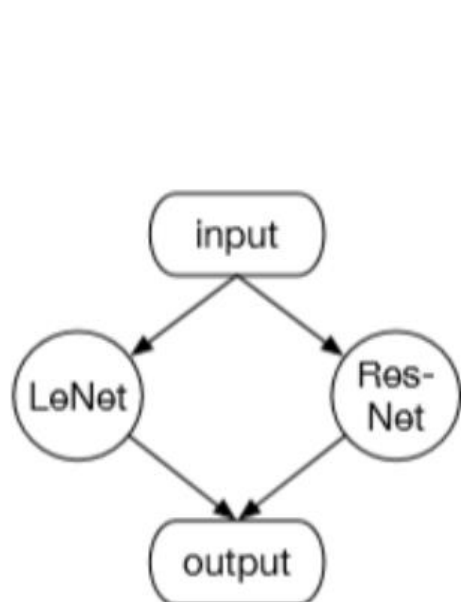


System Framework

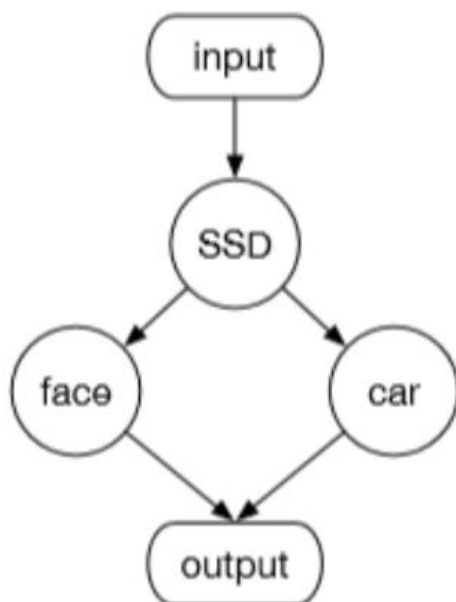


Global Scheduler

- Scheduling Complex Queries

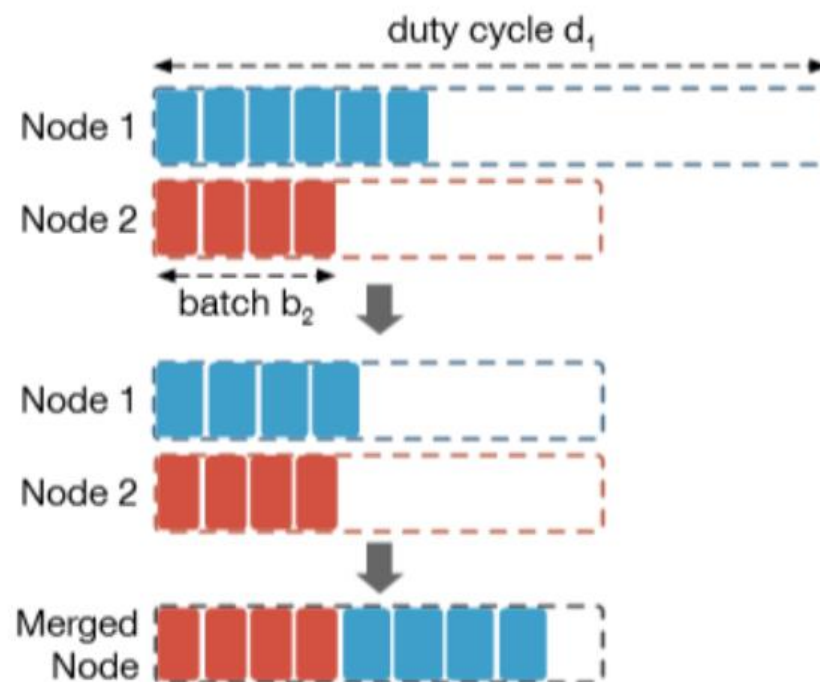


(a) Game



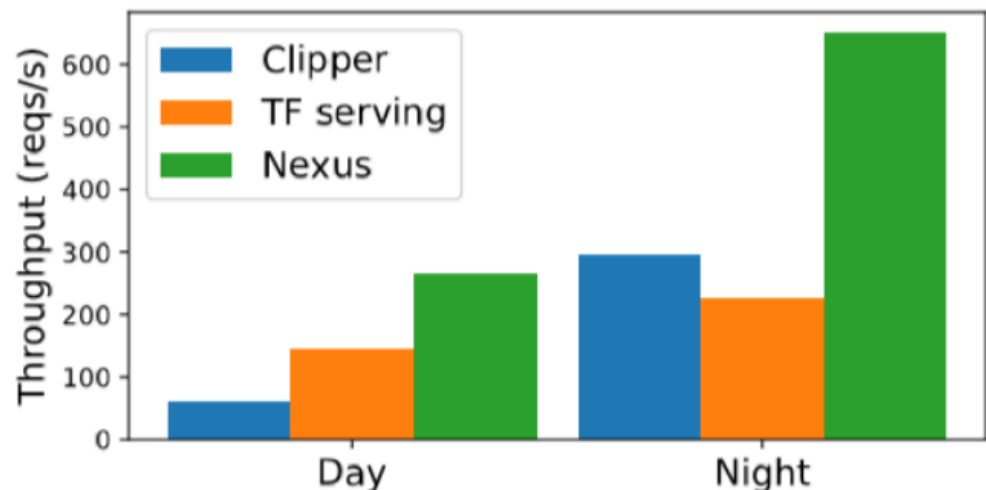
(b) Traffic

- Squishy bin packing

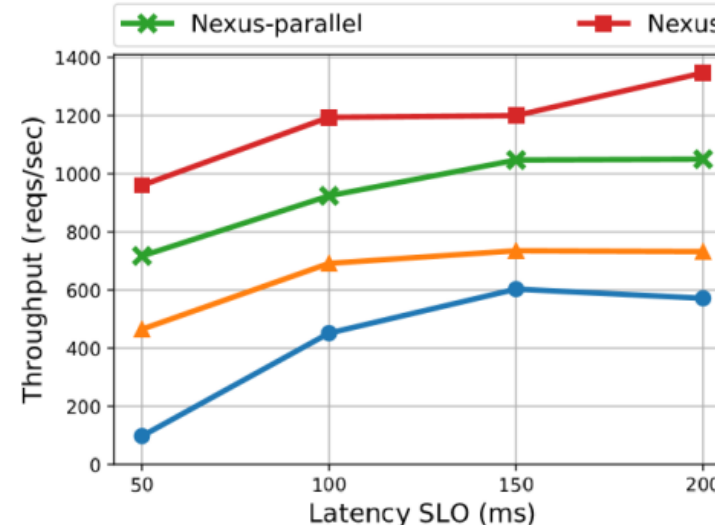
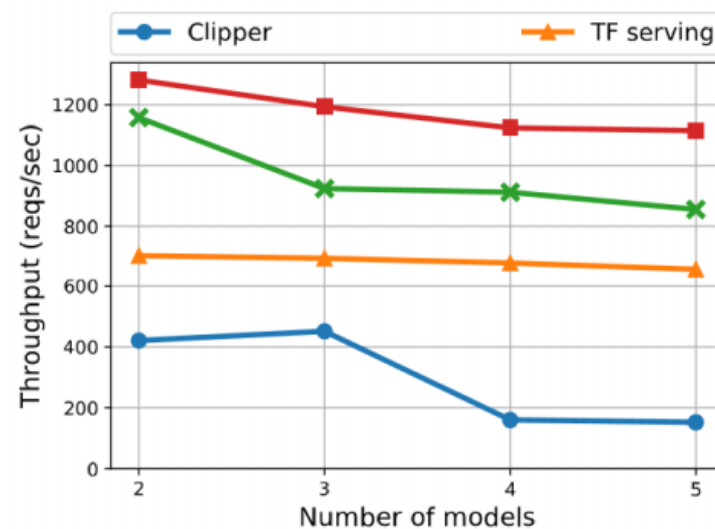


Result

- Compare with Clipper and Tensorflow Serving



Throughput for traffic analysis applications on 16 GPUs.





Q & A