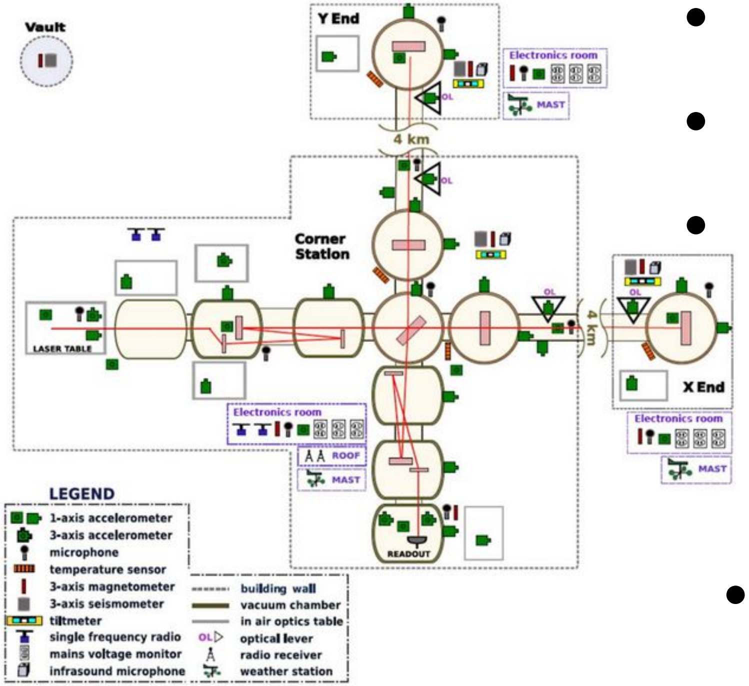
Deep Learning in the Cloud for Gravitational Wave Physics

Alec Gunny[1,*], Dylan Rankin[1], Jeffrey Krupa[1], Saleem Muhammed[2], Tri Nguyen[1], Michael Coughlin[2], Philip Harris[1], Erik Katsavounidis[1], Steven Timm[3], Burt Holzman[3]

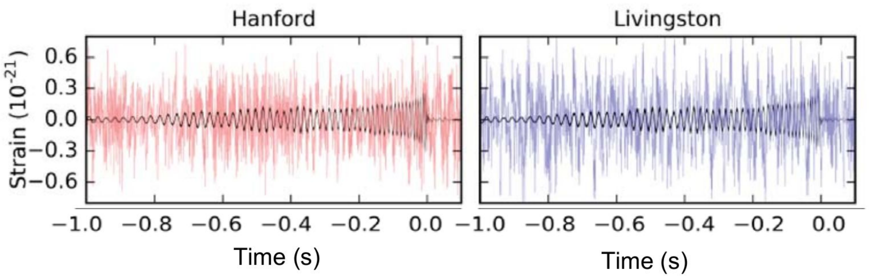
* - Presenter
1 - Massachusetts Institute of Technology
2 - University of Minnesota
3 - Fermi National Accelerator Laboratory

Gravitational Waves and LIGO



LIGO and Virgo Collaborations, *CQG* **33**, 134001 (2016)

- Large scale astrophysical events cause distortions in spacetime known as gravitational waves
- Tiny amplitude of these distortions makes them difficult to detect
- LIGO - pair of enormous interferometers that use destructively interfering lasers to measure perturbations in spacetime



- Measurement of distortion typically given by unitless quantity “strain”, related to relative change in displacement of objects caught in the wave
- Inferred from intensity of photons detected as GWs distort laser paths and bring them in-phase

Noise, MMA, and Real-time requirements

- Environmental noise can degrade the perfect destructive interference of the lasers
- Leads to spurious photon detection, leads to noisy strain measurements
- Makes it difficult to pick out signals with amplitude less than noise, limits detection range
- Auxiliary sensors measure noise for removal

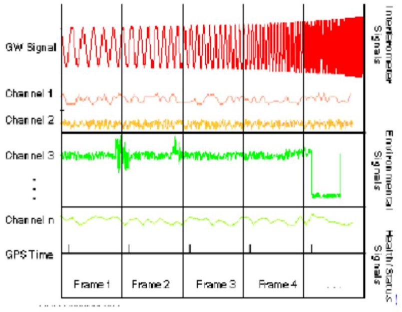
Gravitational-wave Detector Data

Continuous time series (1Hz, 128Hz … 16kHz)

Gravitational Wave channel:
~20GB/day (per instrument)

Physical Environment Monitors (seismometers, accelerometers, magnetometers, microphones etc)

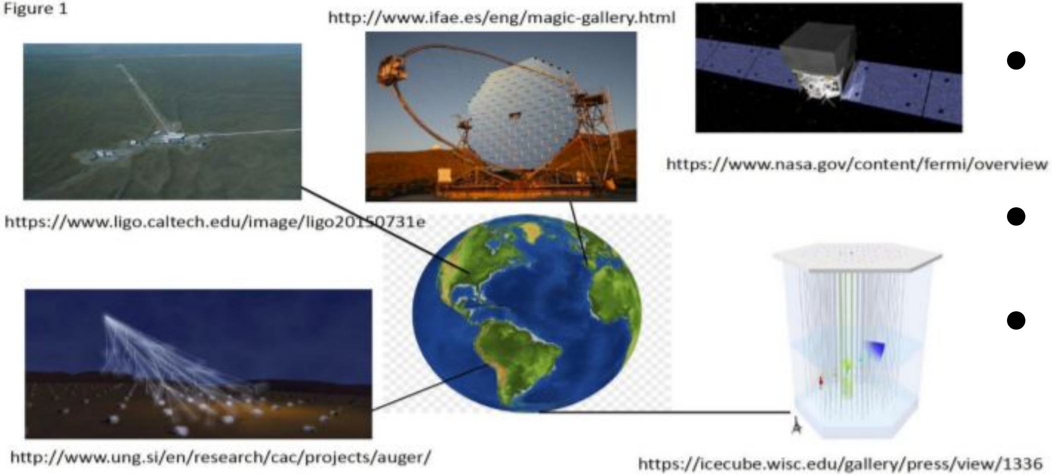
Internal Engineering Monitors (sensing, housekeeping, status etc)



Together with various intermediate data products >2TB/day (per instrument)

Initial and Enhanced LIGO archive (2002-2010) exceeds 1PB of data

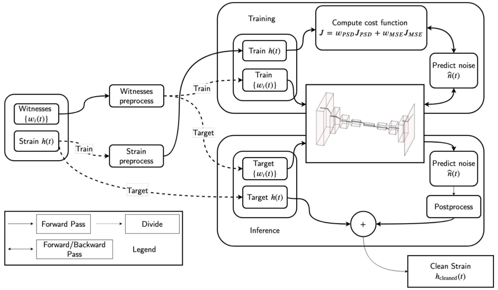
Figure 1

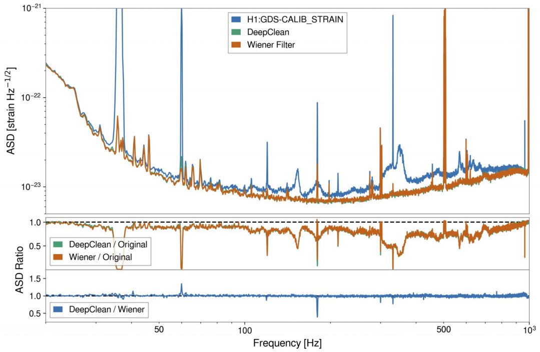


Meszaros et. al. https://arxiv.org/pdf/1906.10212.pdf

- Multi-messenger astrophysics offers promising insights by comparing different cosmic messengers from same phenomena
- LIGO + VIRGO critical for detecting and locating events to alert other observers
- Noise subtraction and downstream algorithms need to work in real-time to capture as much data as possible during O4 data collection

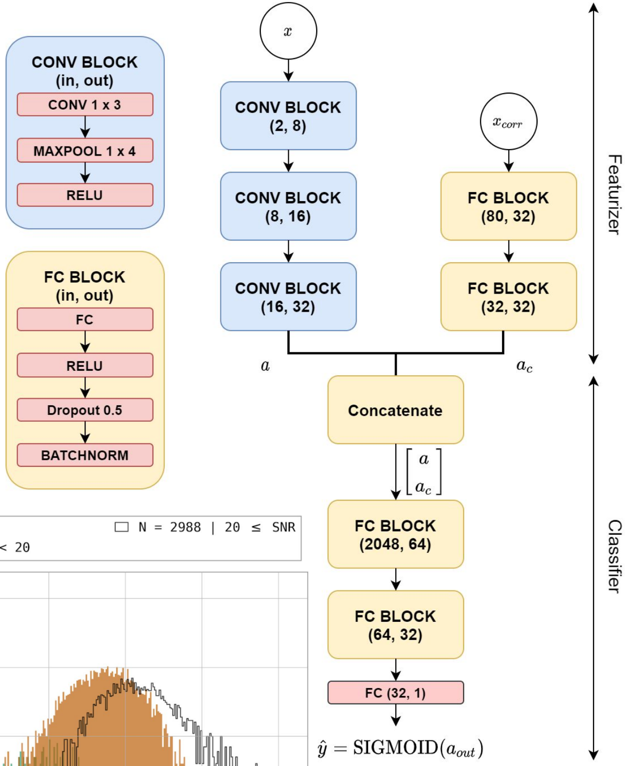
Deep Learning in Gravitational Wave Astronomy

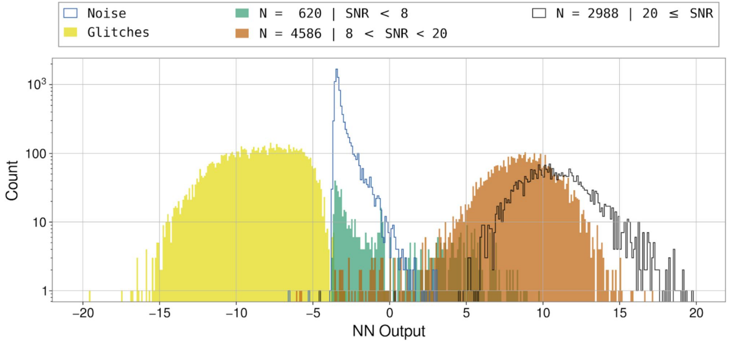




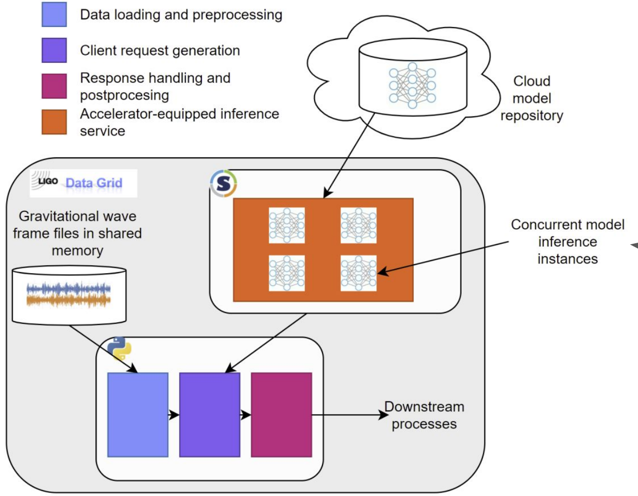
Ormiston, Rich, et al. “Noise Reduction in Gravitational-Wave Data via Deep Learning.” *Physical Review Research*, vol. 2, no. 3, July 2020, p. 033066. *arXiv.org*, doi:10.1103/PhysRevResearch.2.033066.

BBHnet



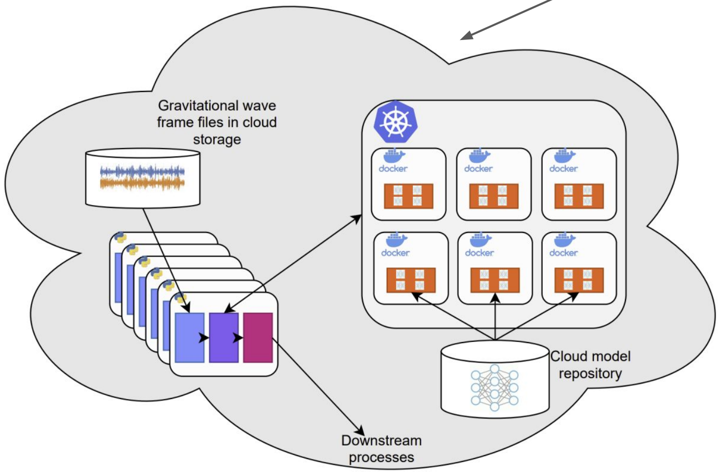


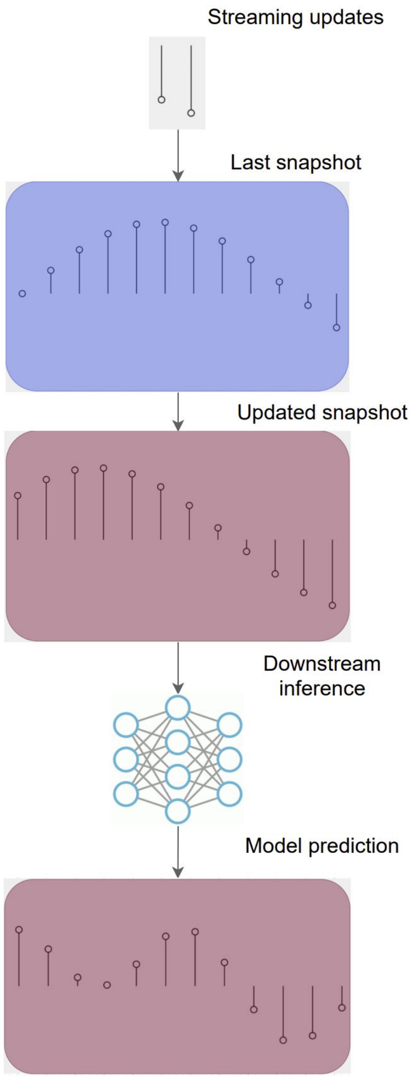
- Presents potential for robust modelling with real-time capabilities
- Investigating potential for arbitrary anomaly detection to learn new physics
- Strong, efficient performance non-trivial to achieve

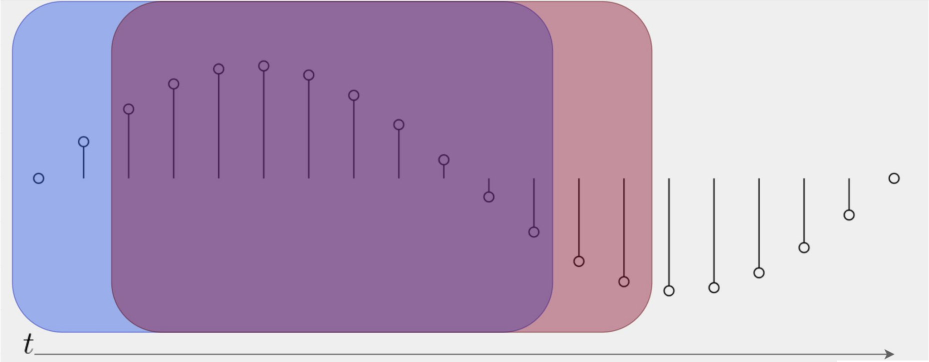


Inference-as-a-Service model enables and simplifies scalable, high-performance DL inference on heterogeneous systems in a variety of deployment scenarios

- Low-latency, "**online**" deployment for data preprocessing and MMA triggering
  - Deploy on local resources to minimize data loading and client-server I/O
- High throughput, "**offline**" deployment for validating new models and searching for new physics
  - Deploy in cloud to take advantage of strong scaling and elastic availability of resources
  - Faster results -> faster iteration -> better ideas

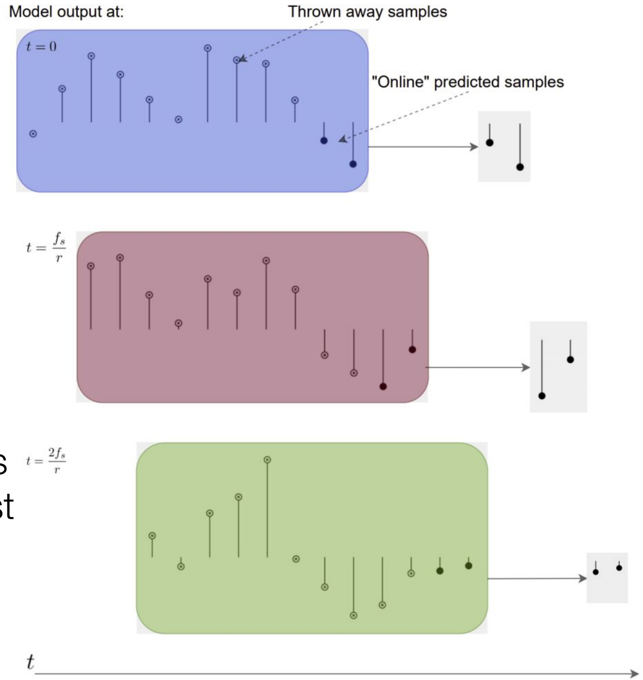






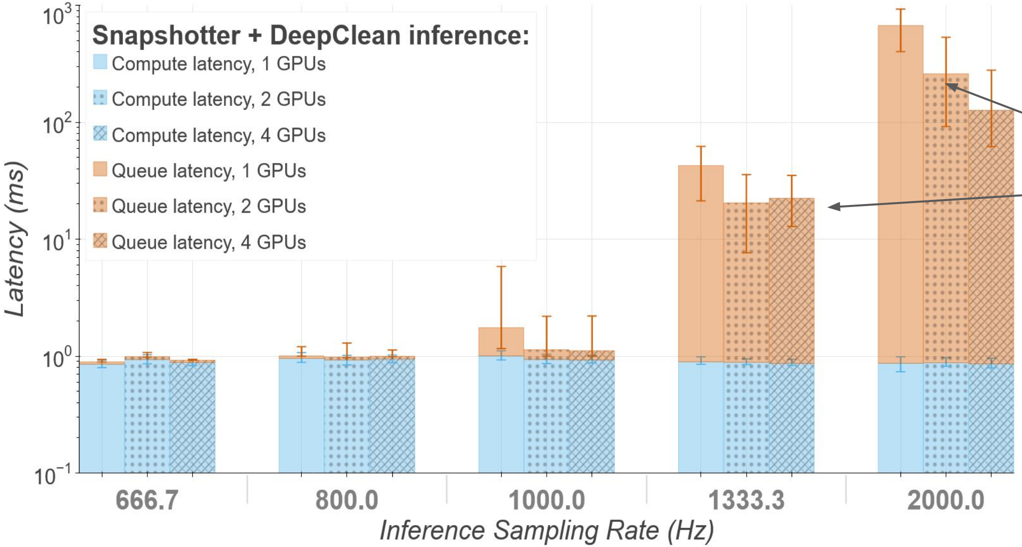
- Models act on fixed length snapshots of timeseries
- Decide what rate to sample snapshots at inference time
- Overlapping data creates I/O bottleneck

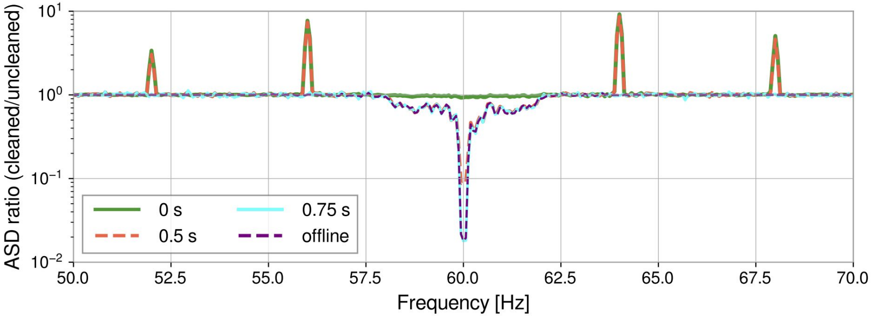
- Built caching model that maintains snapshot state on inference service
- Decreases bandwidth by factor of snapshot sampling rate
- Allows for lower latency and higher quality inference at no extra cost



- Similar overlap for models like DeepClean that output a timeseries
- Adopted “online” strategy for lowest latency
- Potential benefits from aggregation across subsequent outputs

Online data cleaning results





DeepClean capable of removing noise from data in real time, but:

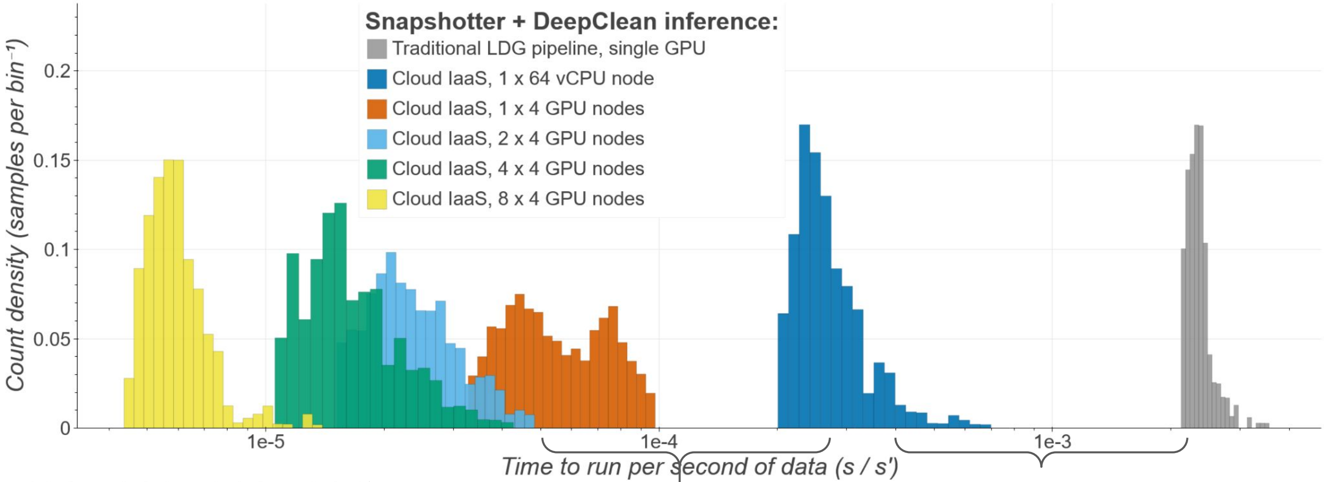
Caching update bottlenecking higher frequency inference

Online output strategy degrading quality of clean

Addressing these issues for O4 run by adjusting training to give better latency-quality tradeoffs at inference time. Measure as part of validation test

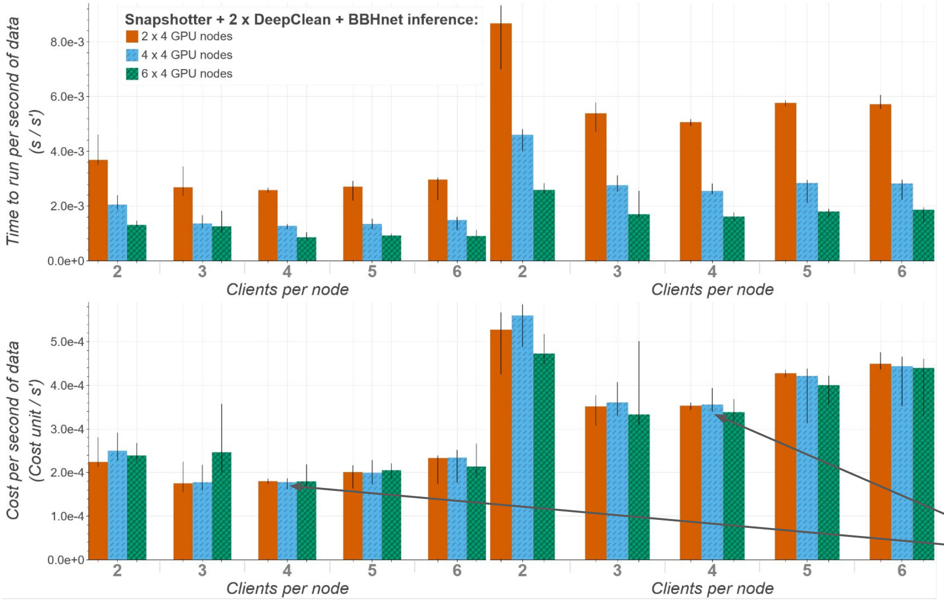
Offline cloud-based results

Large scale processing of data from O2 and O3 runs



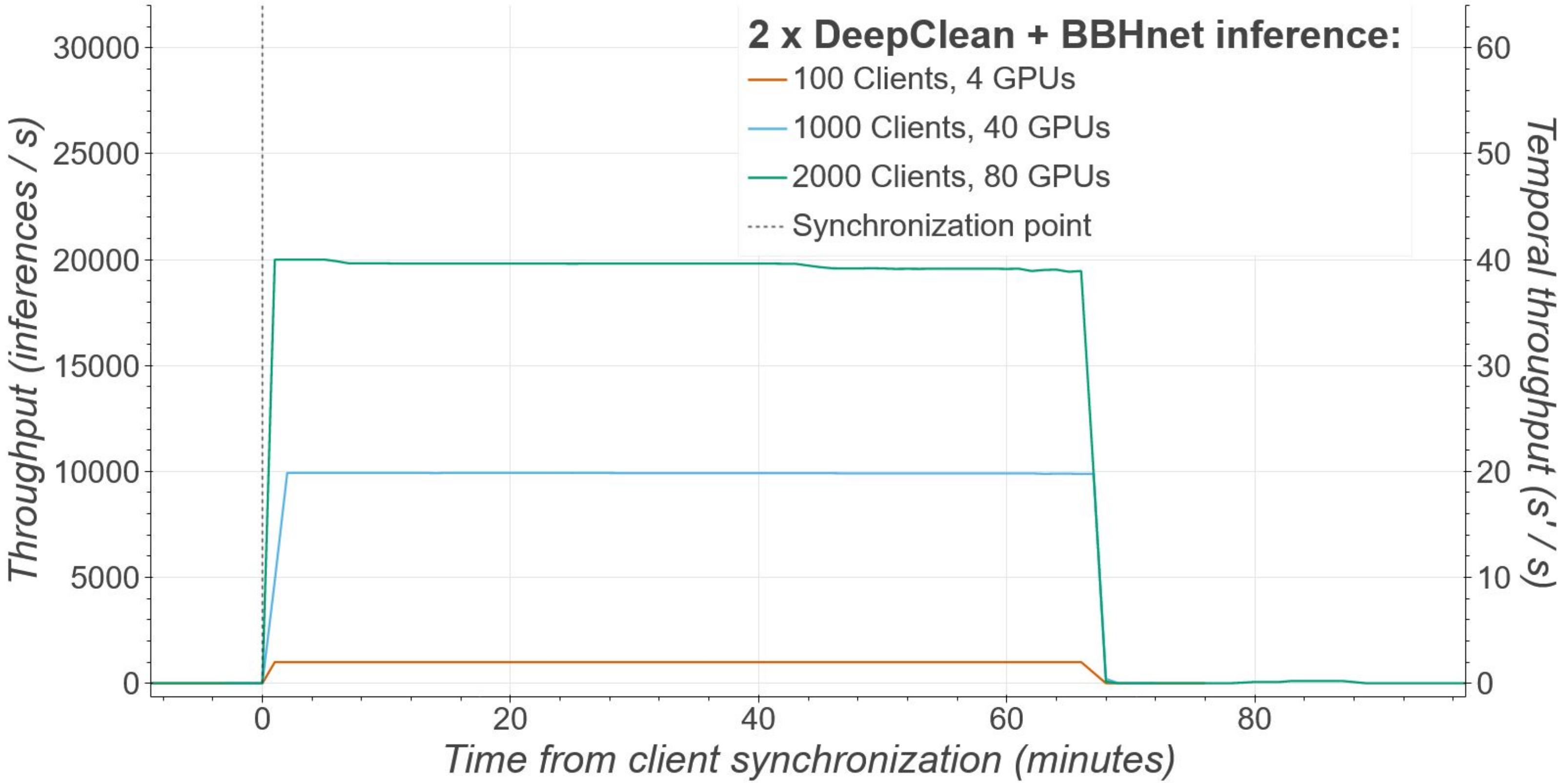
Order of magnitude decrease in processing time from adoption of IaaS model alone

Another factor of 5 by adding GPUs, linear decreases from there



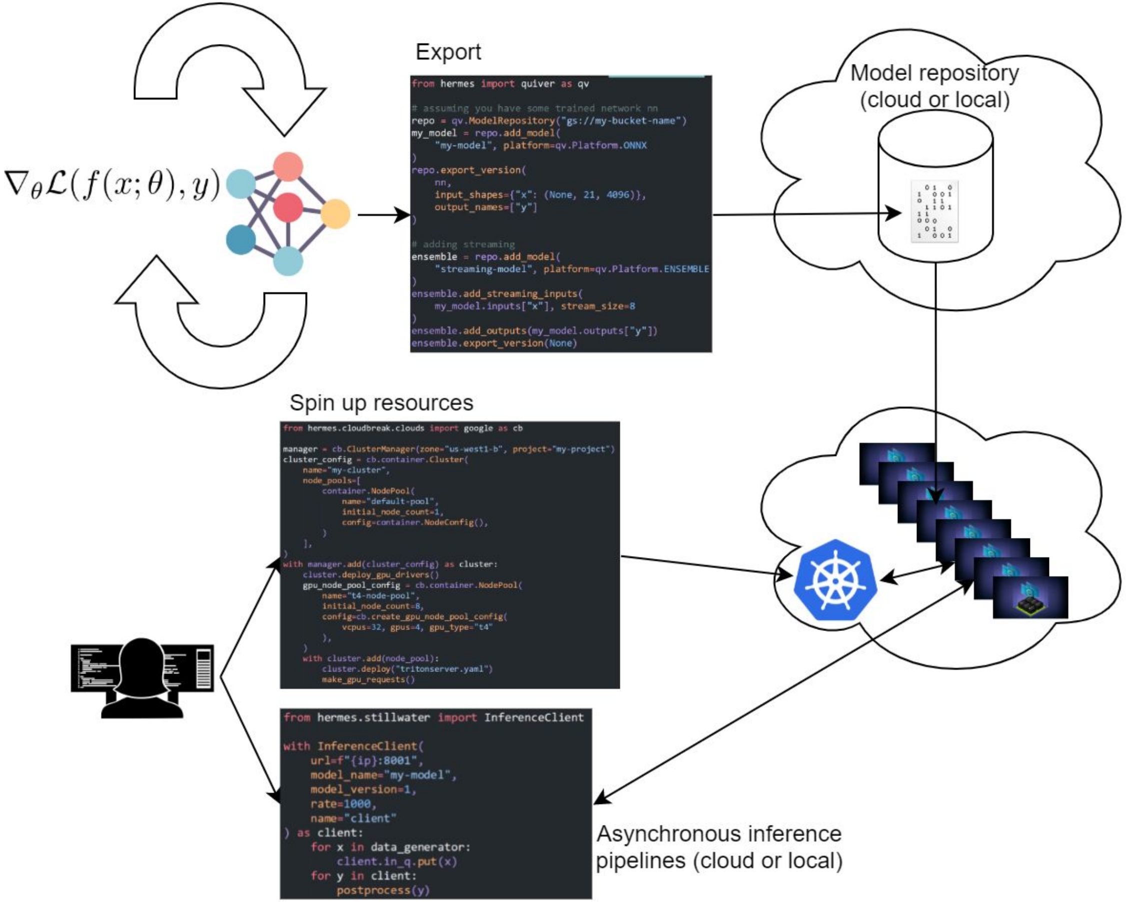
Cost nearly constant as a function of scale

HEPCloud - Scaling up even more



https://github.com/fastmachinelearning/gw-iaas

HERMES: *Heterogeneous-Enabled Real-time Messenger Execution as-a-Service*



Open sourcing end-to-end tools for building portable, robust inference-as-a-service pipelines

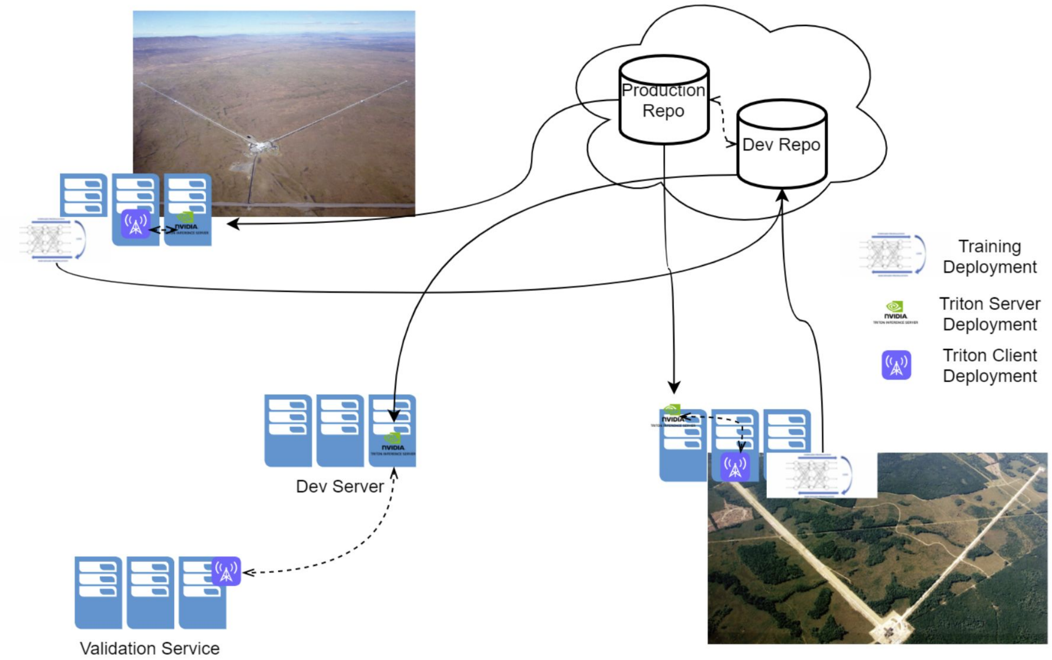
- Agnostic to deployment scenario
- Extensible
- Work-in-progress, welcoming development help!

Next steps

Short term:

- Researching and validating fixes to issues with online DeepClean deployment on large amounts of data
- Using BBHnet and anomaly detection methods to search for new events in O2 and O3 data

Longer term:



Production deployment of online DeepClean pipeline for O4 data collection

- Automated training and validation pipelines deployed on cloud

Thank you!