- Simulation requires little input (mainly beam flux files and photon libraries), has a large memory footprint, uses significant CPU resources and writes back a few large files.
- Data reconstruction reads in large files, has an intermediate memory footprint and uses $\sim$ 10 sec/MB of input data.
- Data reduction reads the reconstructed data and produces small tuple outputs for further analysis. Reduction uses ~ 0.1 sec/MB of input data and is generally IO limited.
- Data analysis consists of repeated access to smaller tuple outputs for calibration and parameter estimation.

Each of these use cases is best suited for a different combination of data/cpu resources and the global compute model should be able to allocate resources appropriately.

We have used the SAM instrumentation to measure the xrootd streaming performance for disk/cpu location combinations.

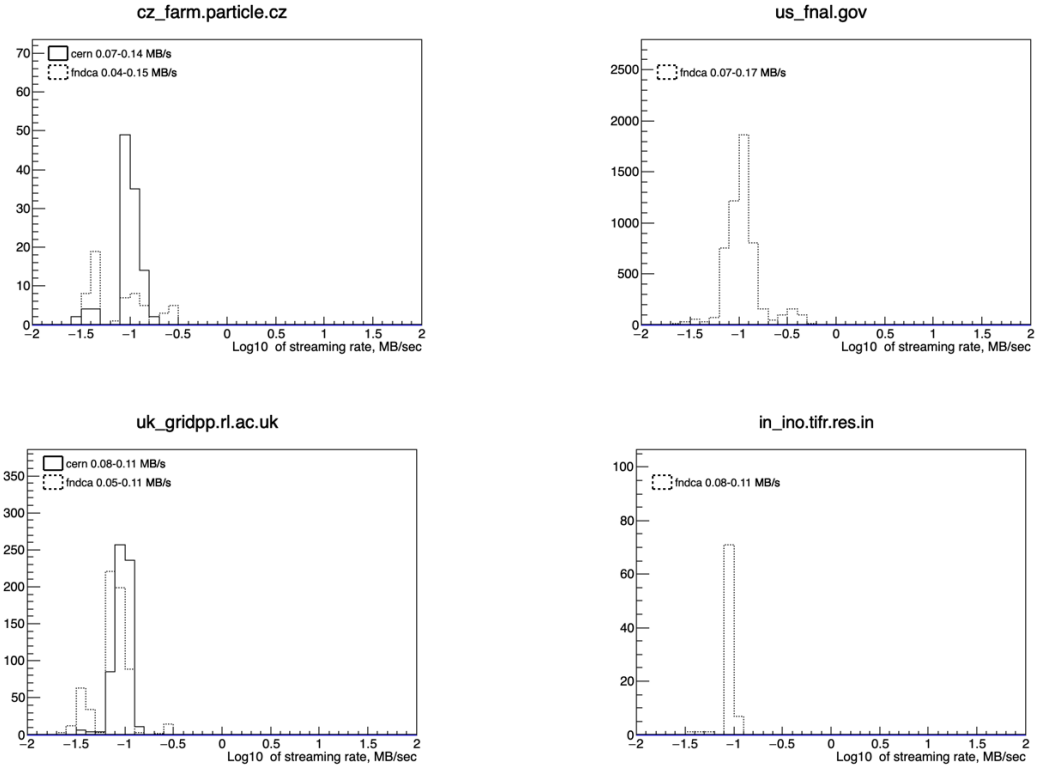

Figure 1.1: Streaming speeds for reconstruction jobs running at different locations. Raw data are stored at CERN and Fermilab. The histograms show the log10 of the inferred streaming rate (wall time/file size) for reconstruction jobs running at selected sites with different data sources.

1.2.1 Implications for data and processing placement

The current study indicates that for CPU dominated applications, notably reconstruction of raw data and simulation, the relative location of CPU resources is not critical. Wall time/MB is similar regardless of location.

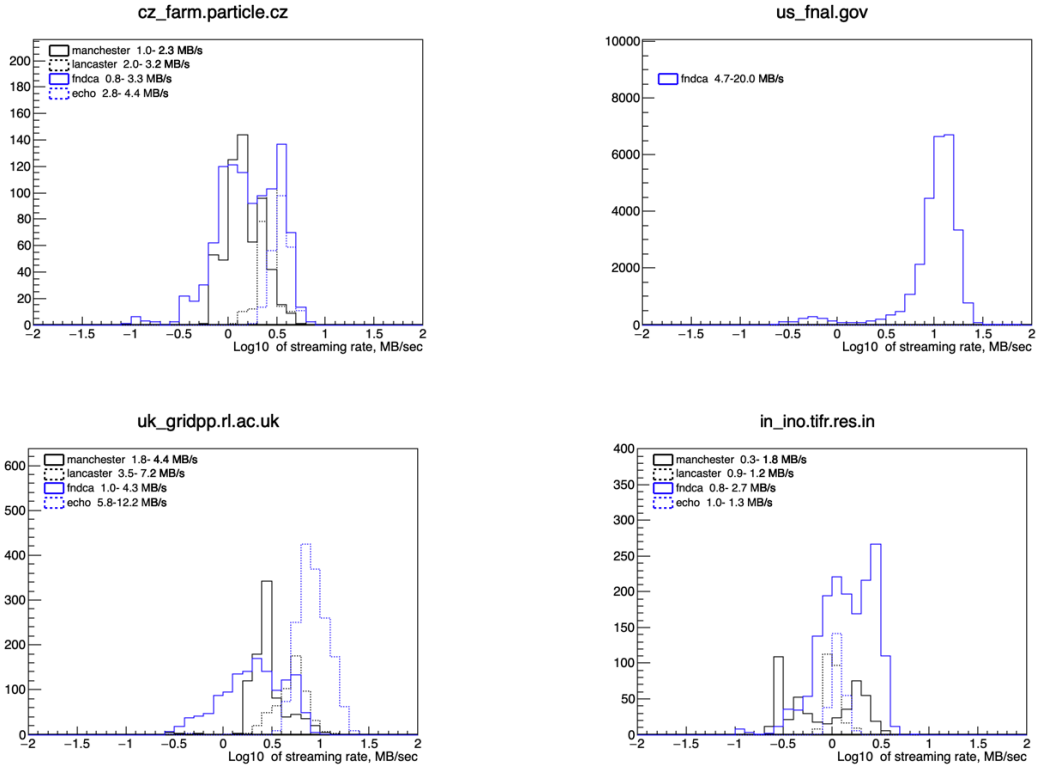

Figure 1.2: Streaming speeds for tuple creation jobs running at selected locations during a test in early January 2022. Reconstructed simulation files were located at sites in the UK and at Fermilab. The histograms show the log10 of the inferred streaming rate (wall time/file size) for tuple creation jobs running at selected sites with different data sources.

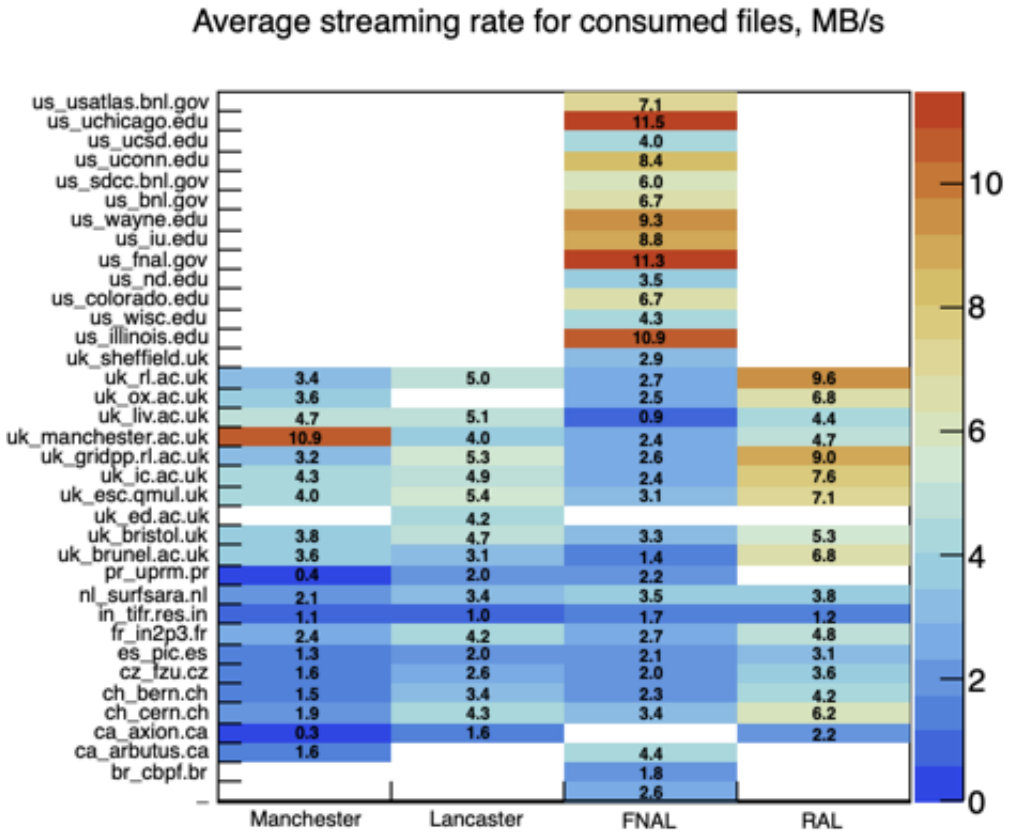

Figure 1.3: Streaming speeds for tuple creation jobs running at multiple locations during a test in early January 2022. Reconstructed simulation was stored at sites in the UK and at Fermilab. The average estimated streaming rates are plotted as a function of disk location (x-axis) and compute site (y-axis). Jobs in the US were required to use Fermilab disk but international sites were tested with multiple samples.