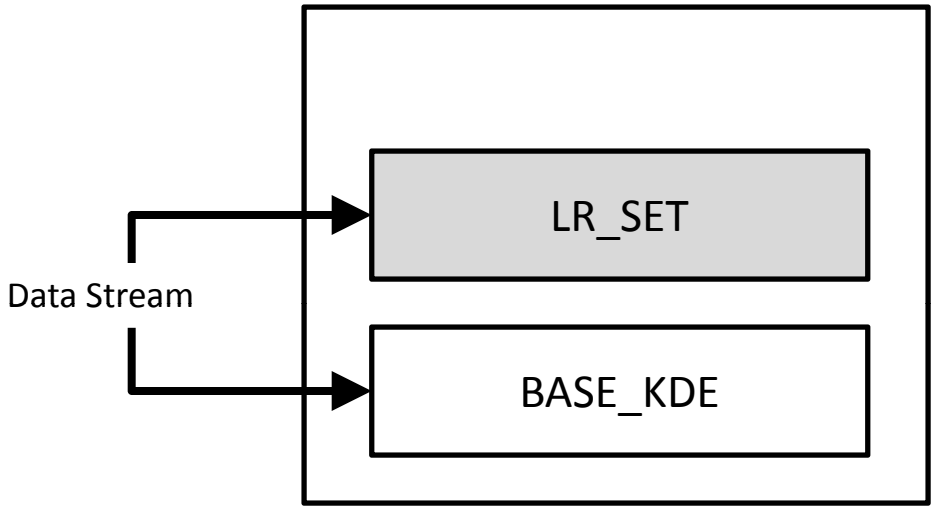
Local Region KDE – Update 2

Arnold Boedihardjo

LR Integration into any existing BASE_KDE



- Construct LR as data sample arrives in tandem with the BASE_KDE
- Employ incremental clustering that minimizes SSE
  - Set |LR_SET| to some max number

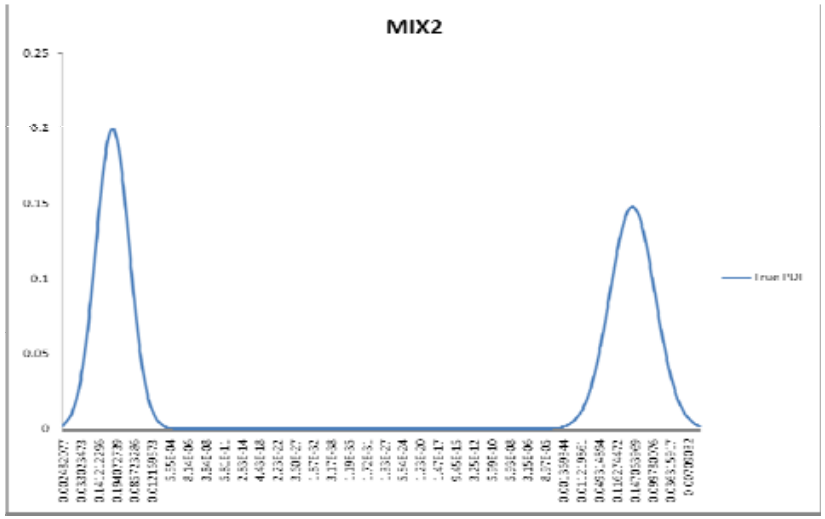
Automatic Local Region Number Determination

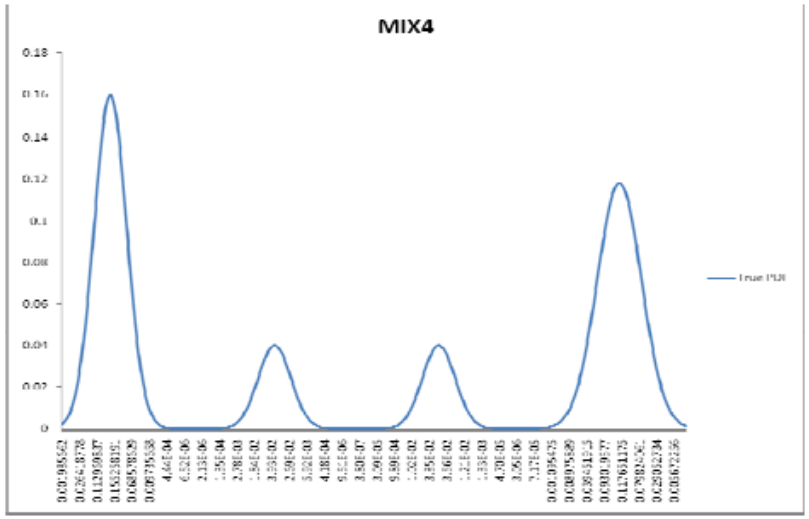
- Perform agglomerative clustering on LR_SET at query-time
  - For each pair of nearest neighbor LRs, merge the pair if $STDEV(LR_{merge}) \leq STDEV(LR_1) + STDEV(LR_2)$
- To speed-up the search of nearest neighbors
  - Maintain a priority queue of LR-pair distances
  - Remove top of priority queue (nearest pair)
  - Merge the pair and insert two (potential) merge-able pairs into queue

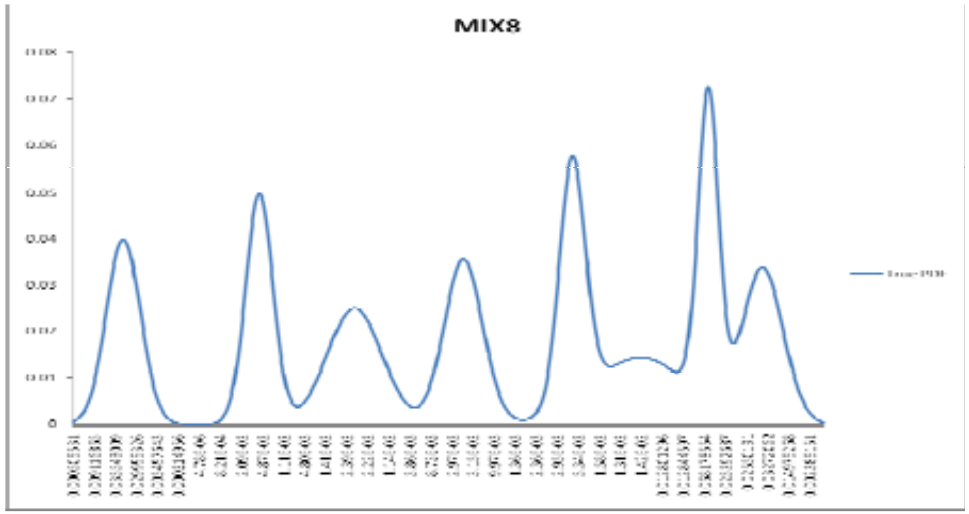
Experiment

- Data sets
  - 3 distributions based on mixture of Gaussians
    - 5 sets of data sets for each distribution (15 data sets total)
  - 2 real world data sets (TRAFFIC, POWER)
- Data stream size = 25K
- Performance measures
  - Estimation quality: integrated absolute error (i.e., area bounded by the estimated and true density)
  - Insertion throughput (samples/time)
  - Evaluation throughput (queries/time)
- Algorithms
  - Cluster (Heinz) KDE, Epanechnikov Time Sampling KDE, Gaussian Time Sampling KDE
  - Cluster KDE+LR, E.T. KDE+LR, G.T. KDE+LR
  - Adaptive KDE
  - Equiwidth Histogram

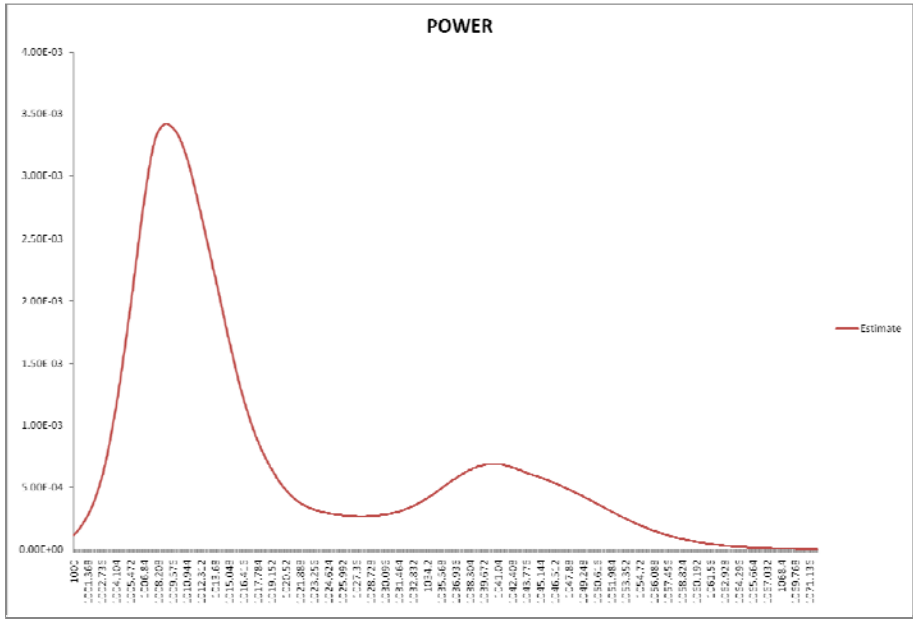
Simulated distributions

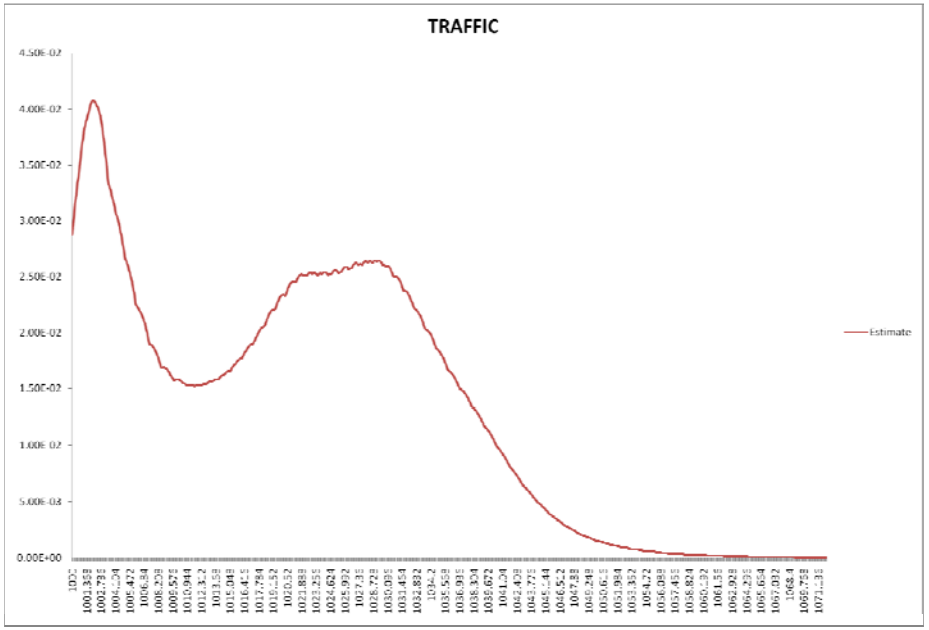






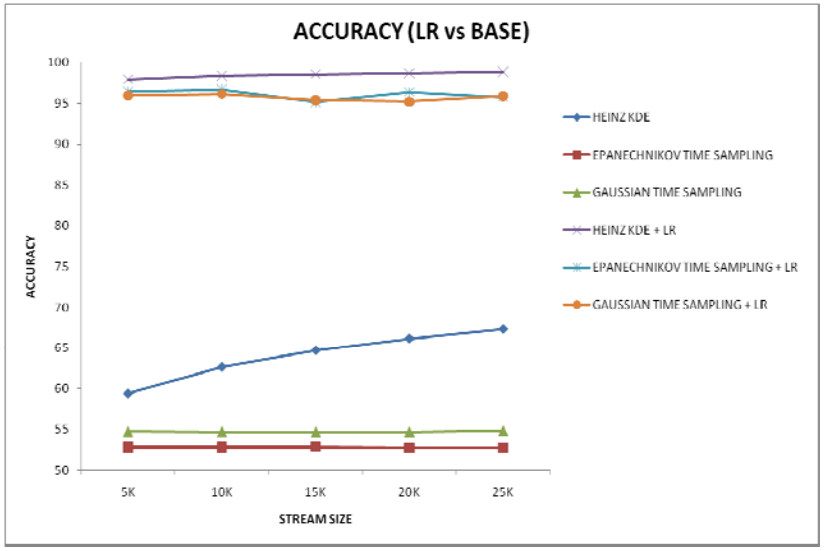
Real-world Distributions (Full AKDE)

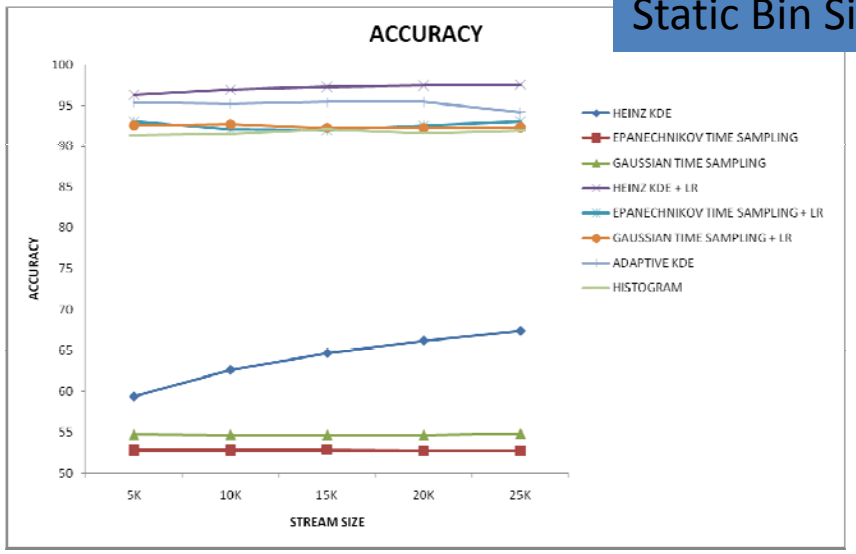


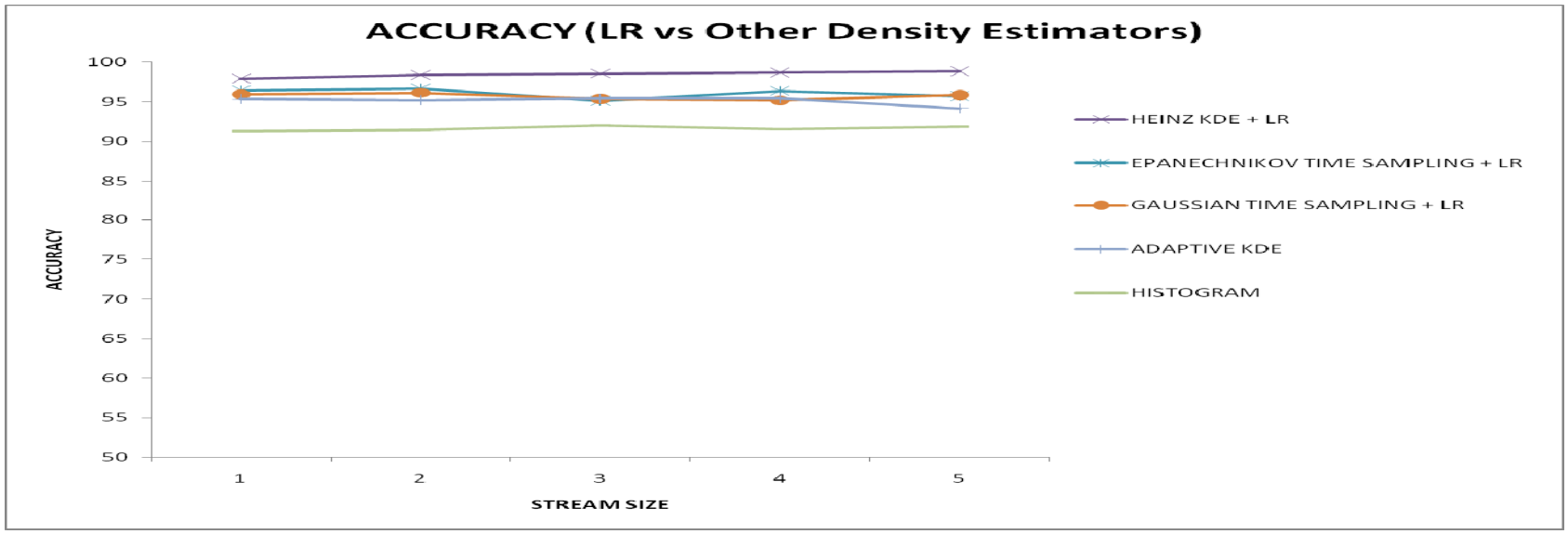


Estimation Quality (MIX2) (STDEV < 1.5%)

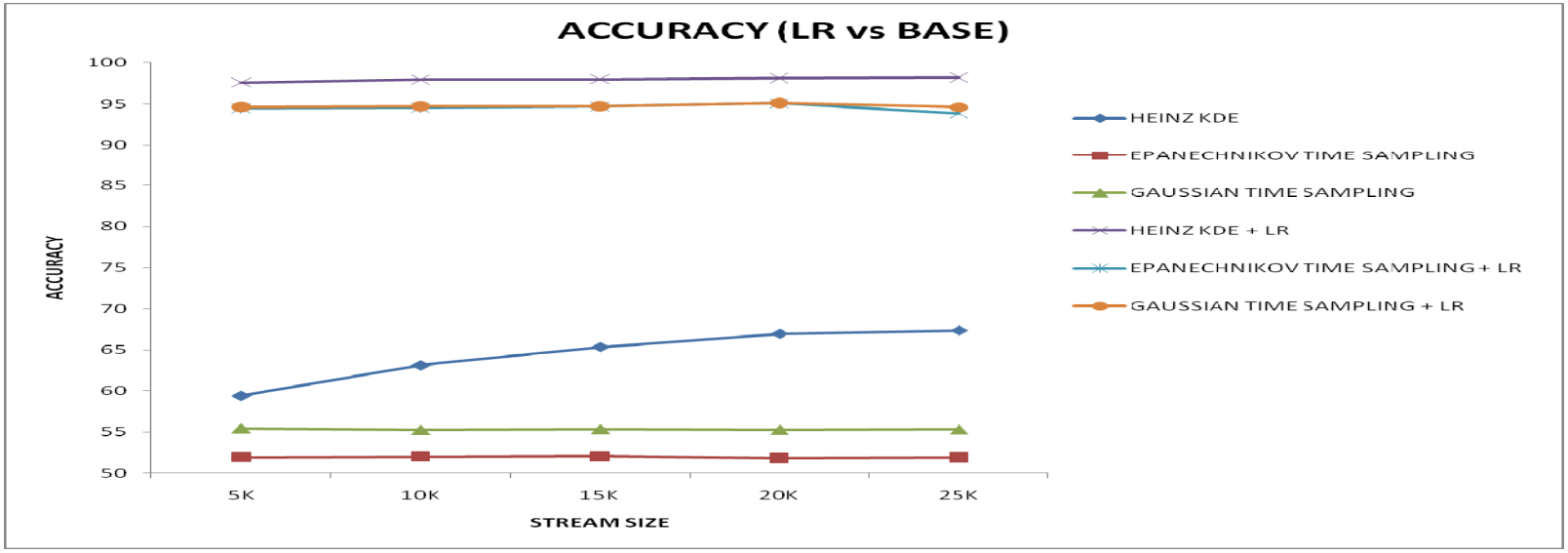
Static Bin Size

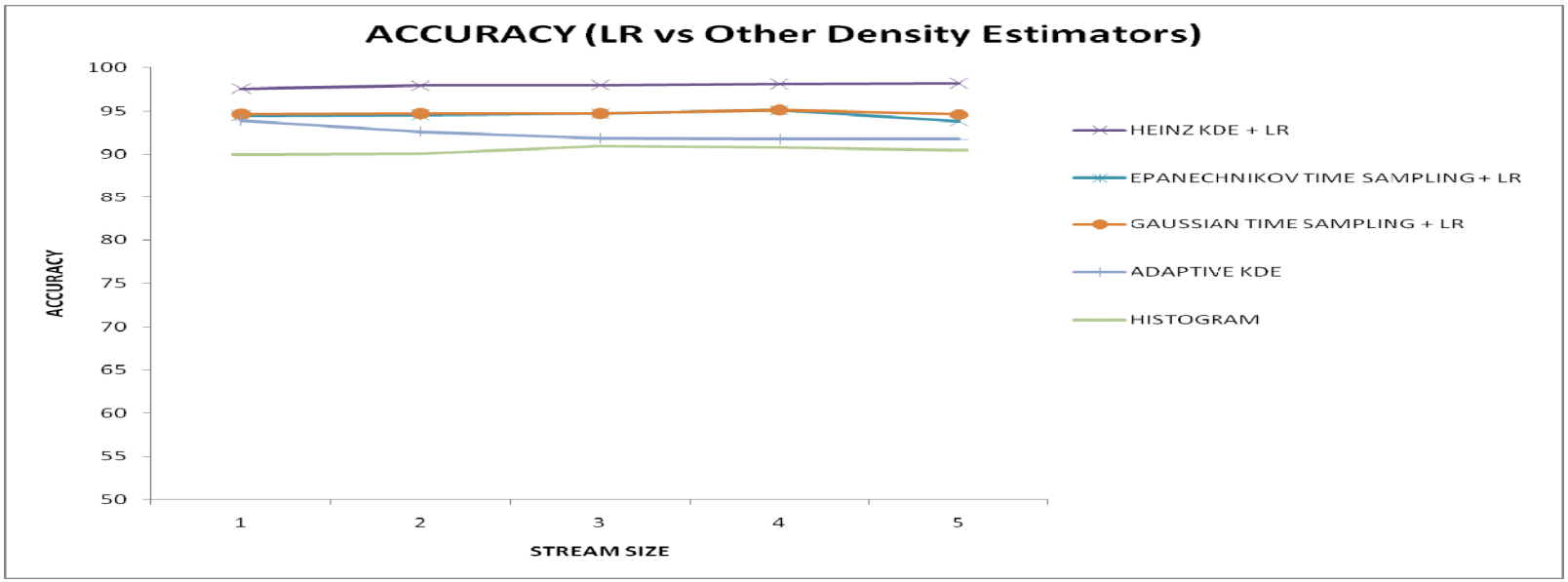




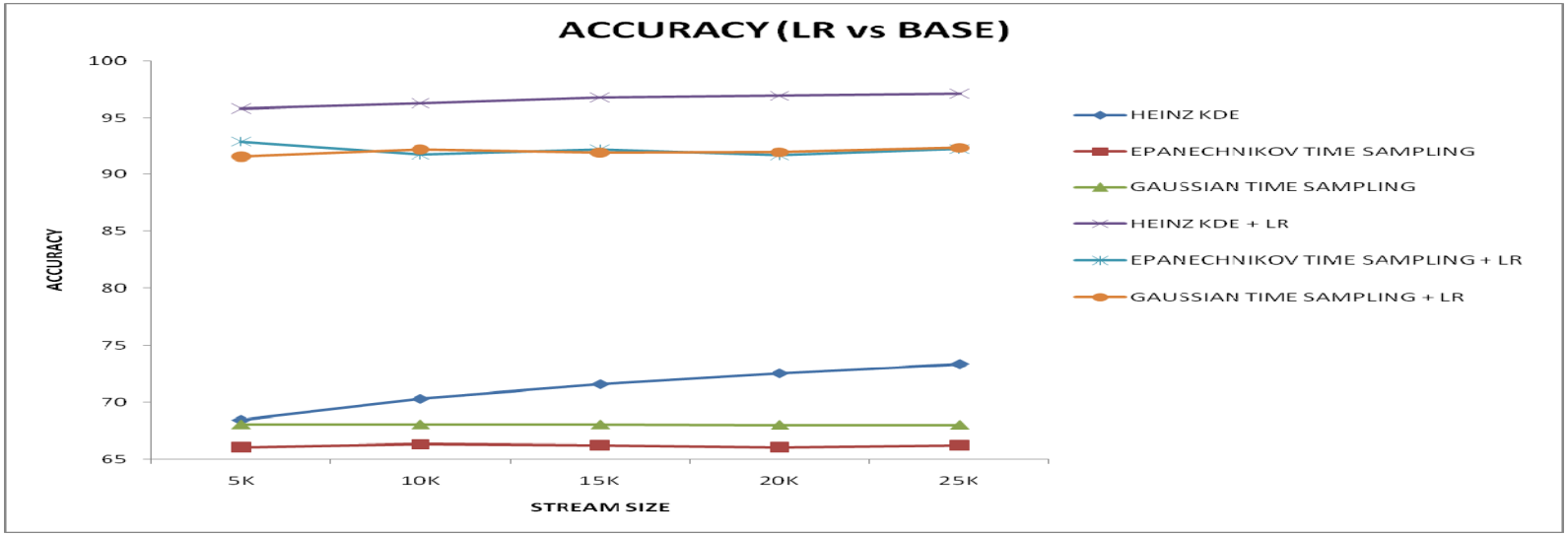


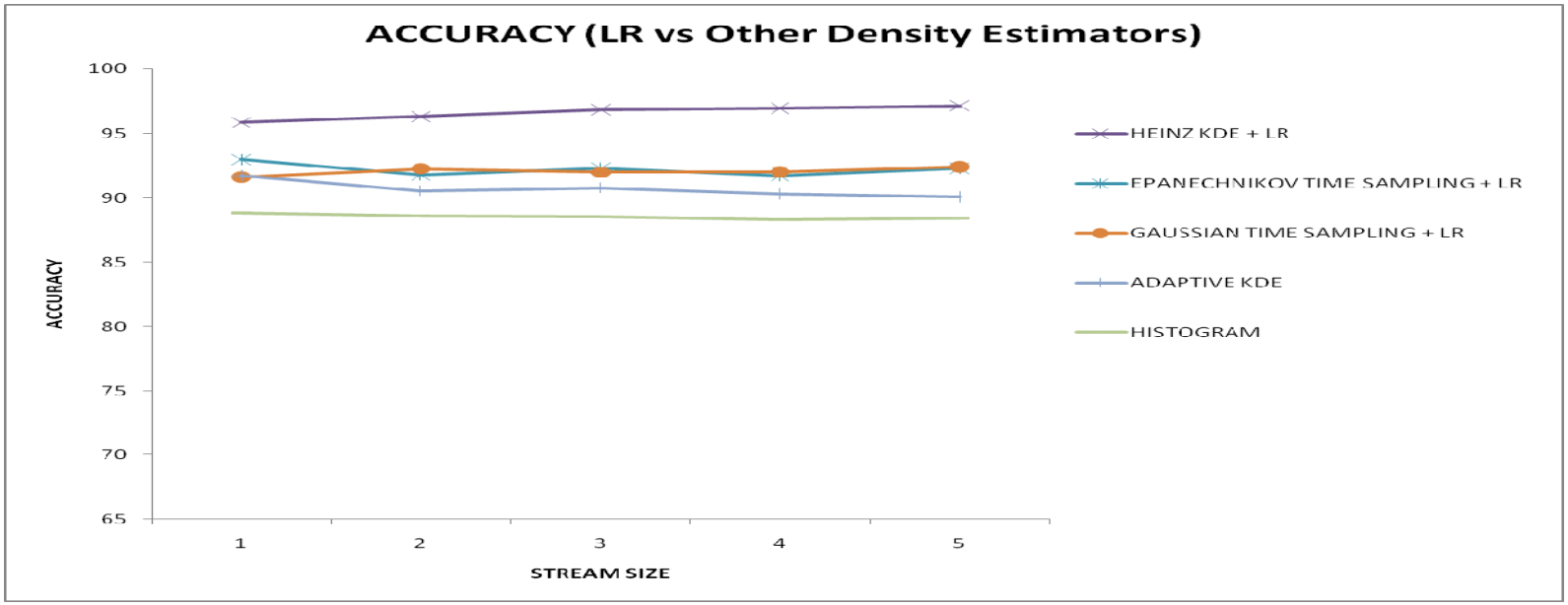
Estimation Quality (MIX4) (STDEV < 1.5%)



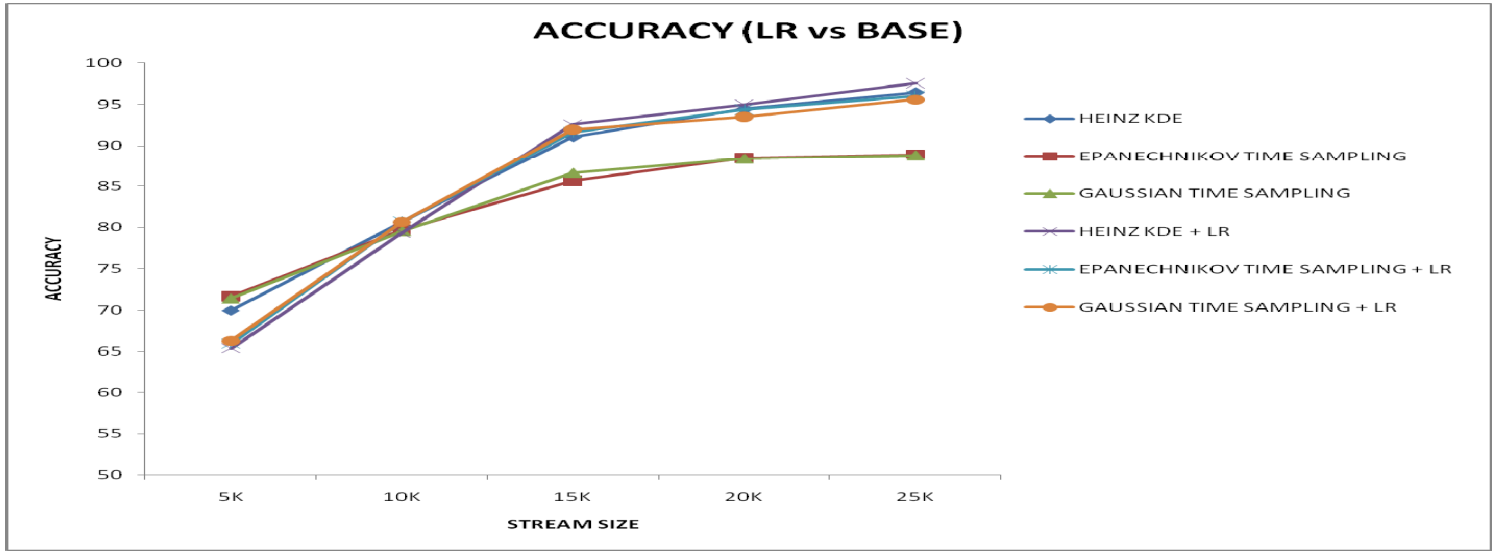


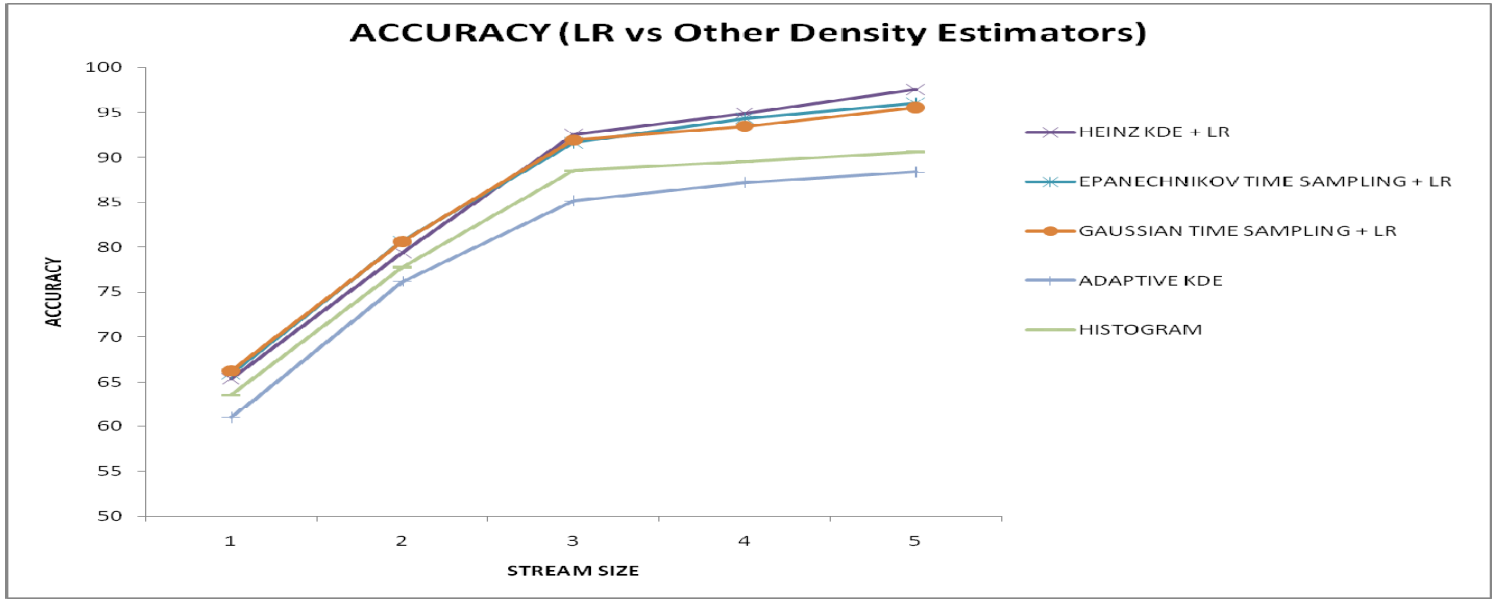
Estimation Quality (MIX8) (STDEV < 1.5%)



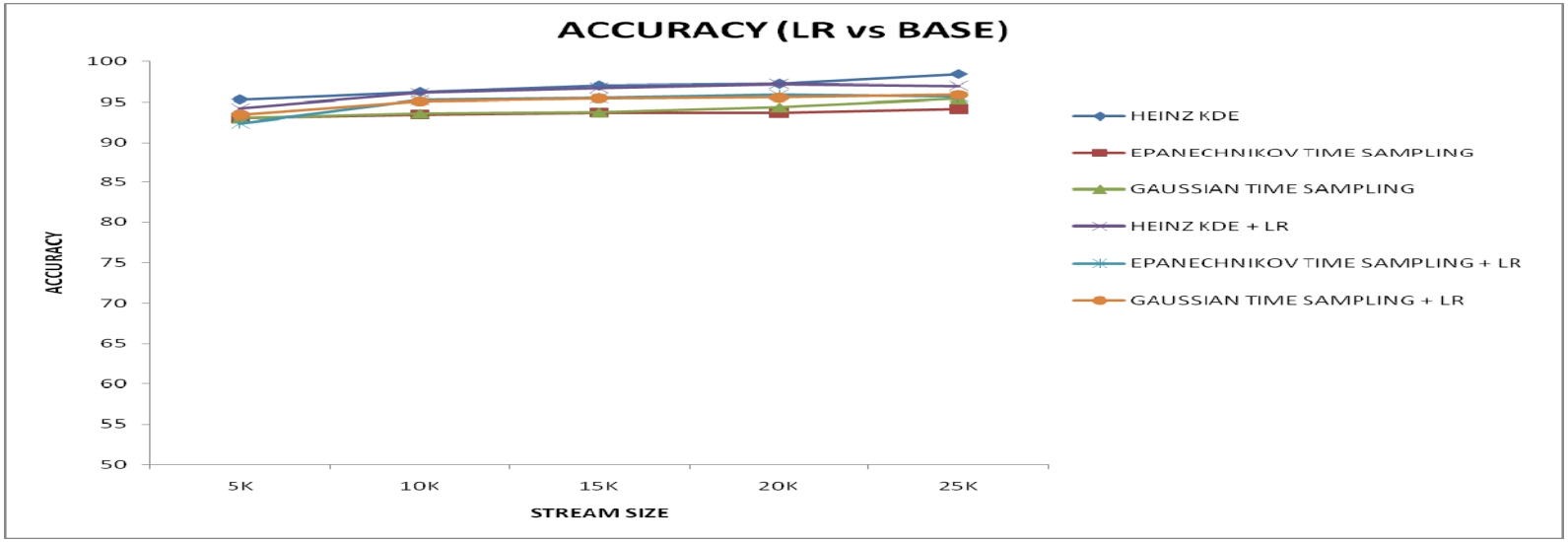


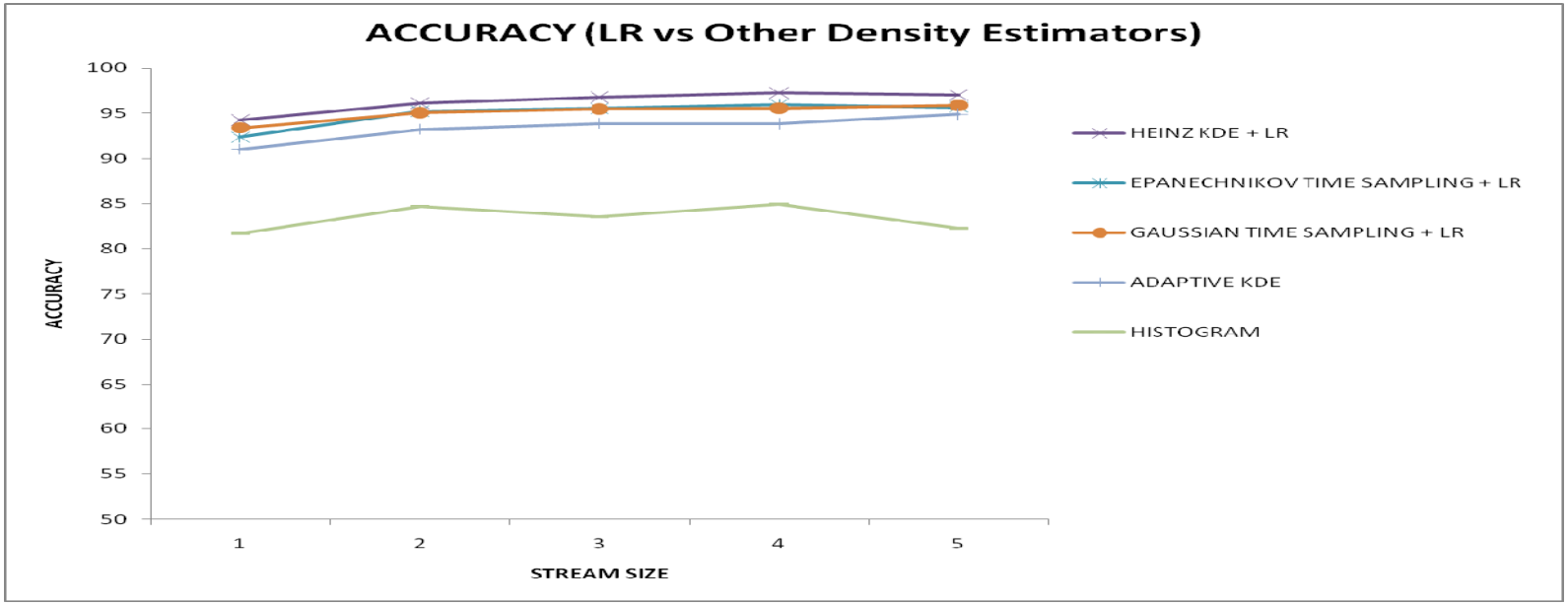
Estimation Quality (POWER) (STDEV < 1.5 %)



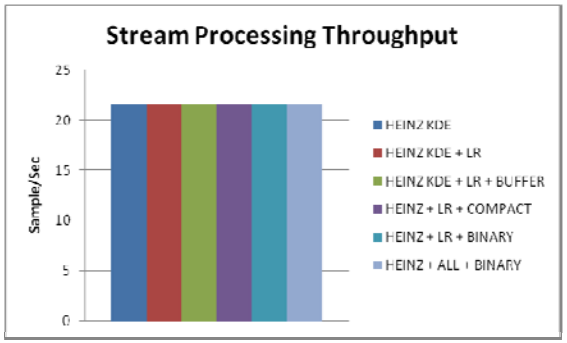


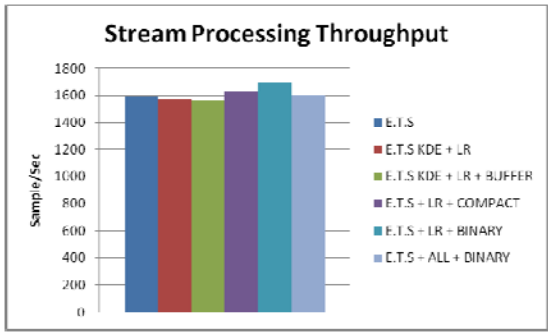
Estimation Quality (TRAFFIC) (STDEV < 2.0%)

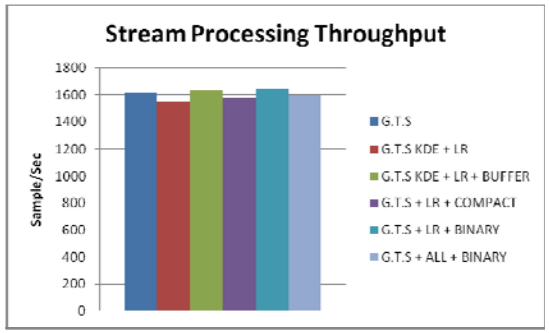


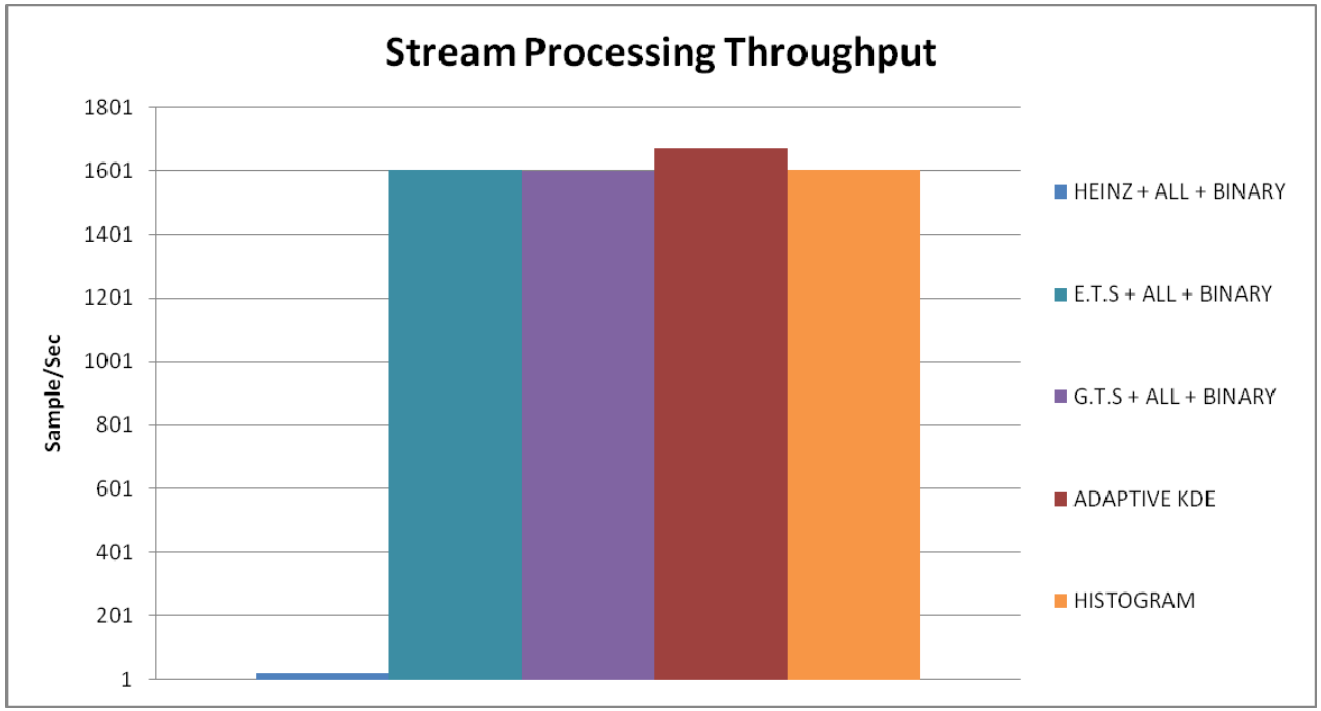


Stream Processing Throughput (MIX2, MIX4, MIX8)

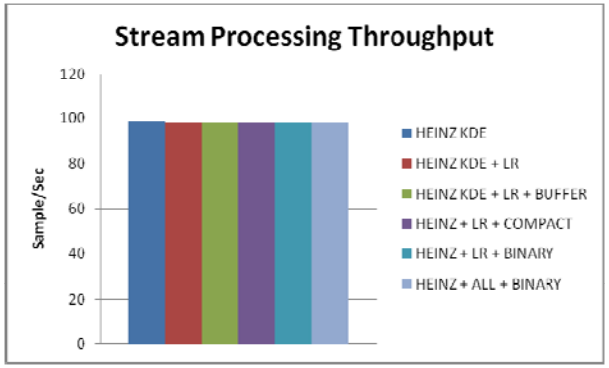


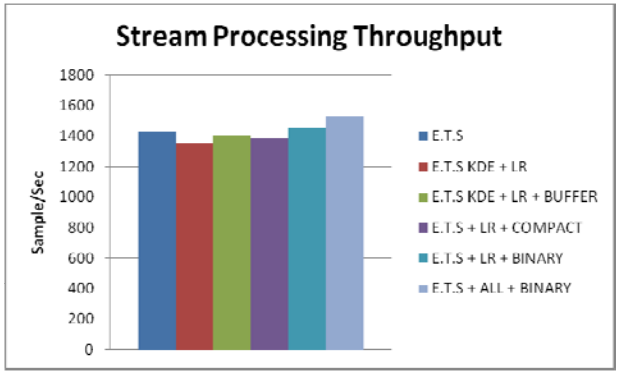


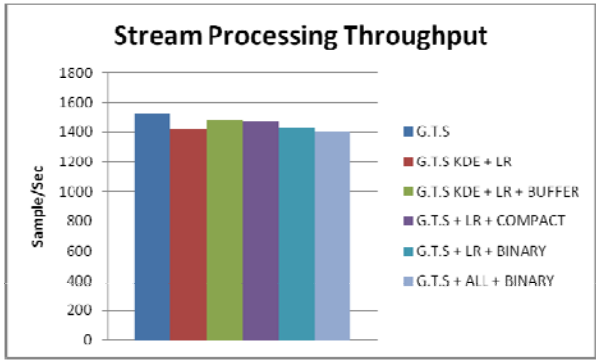


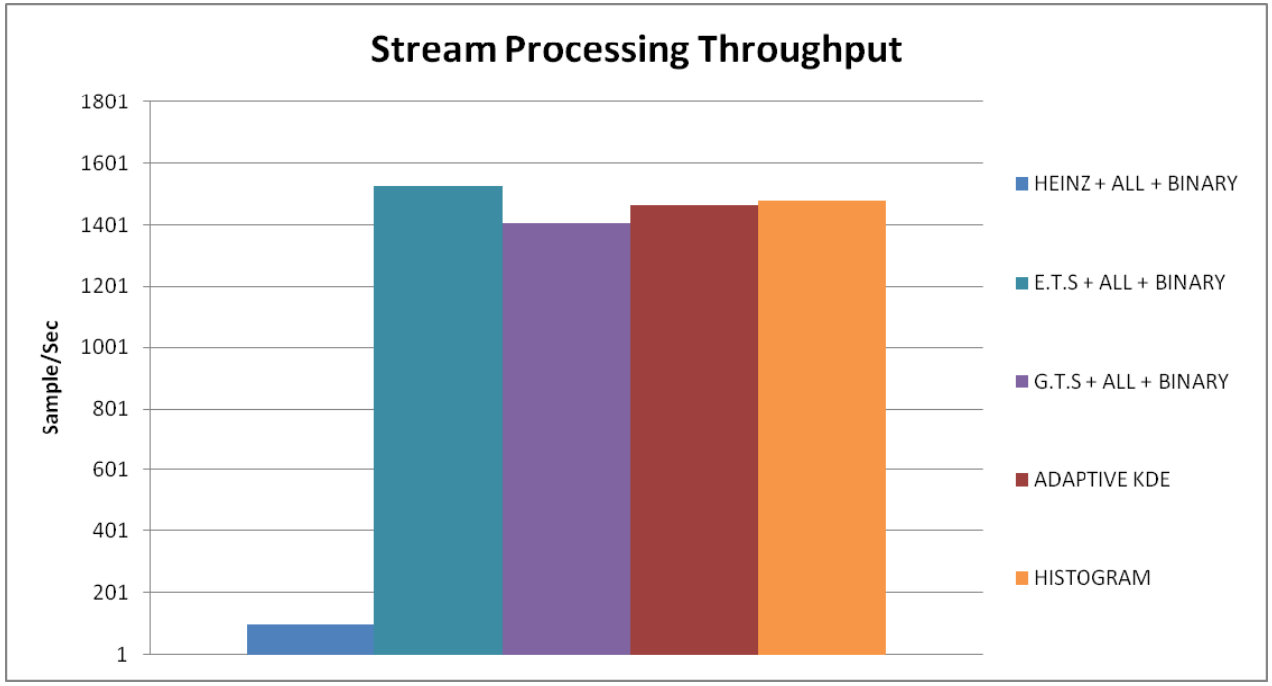


Stream Processing Throughput (POWER)

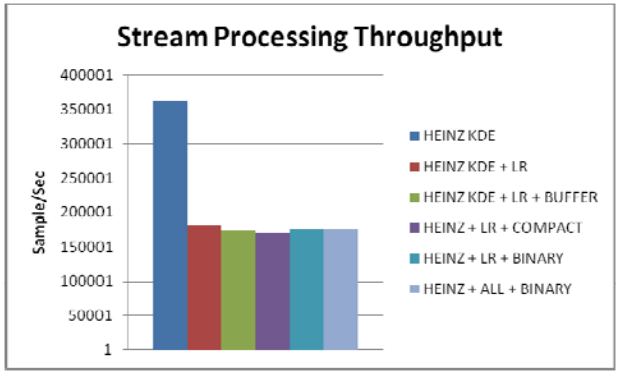


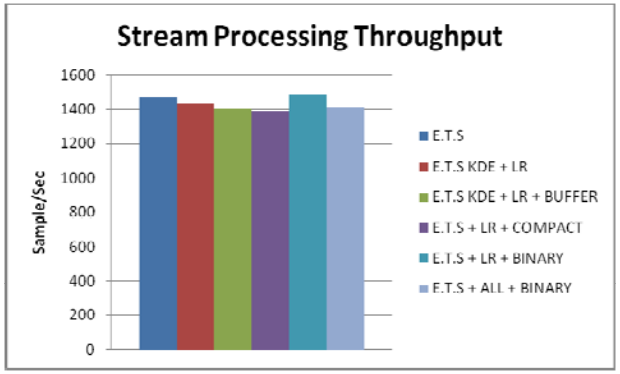


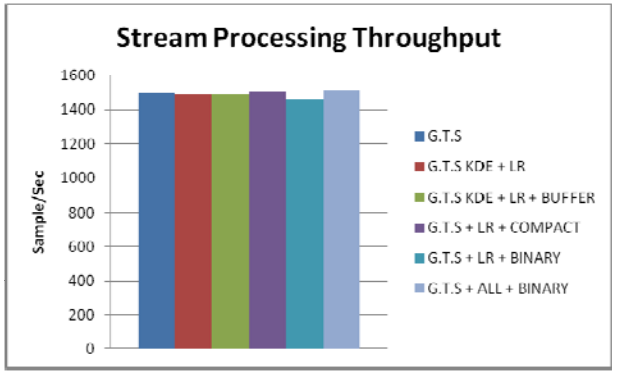


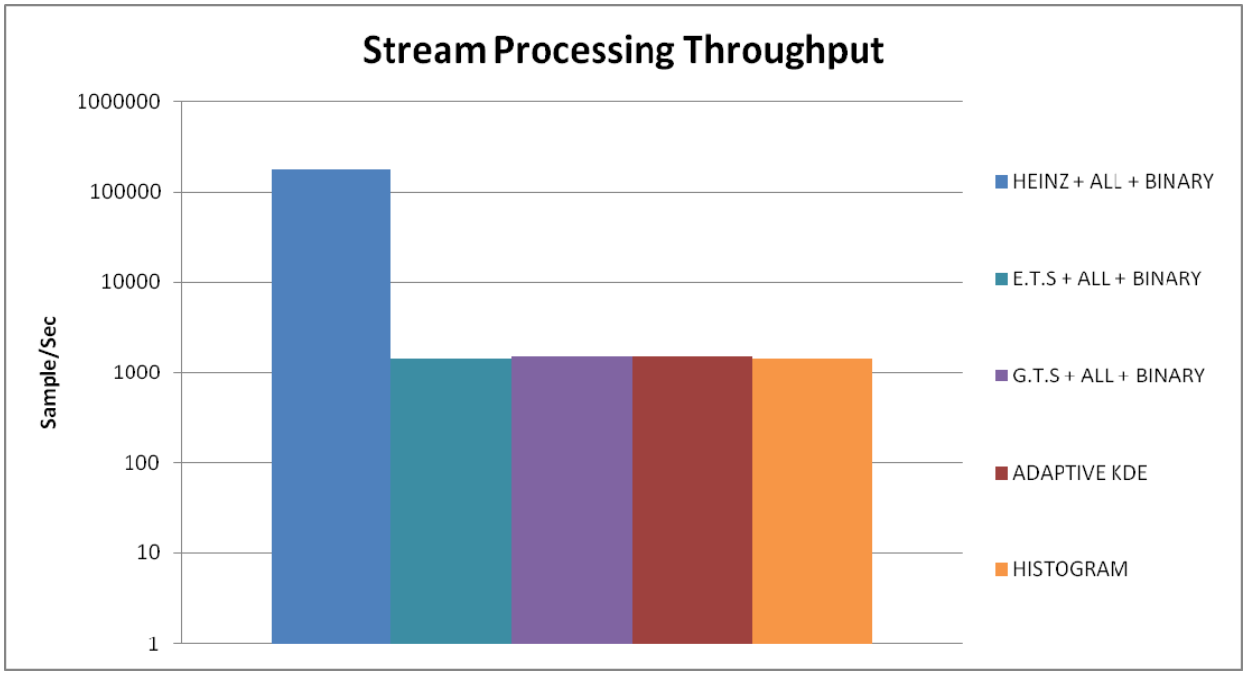


Stream Processing Throughput (TRAFFIC)

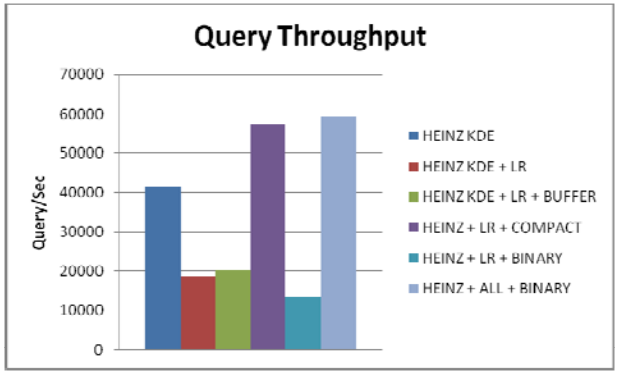


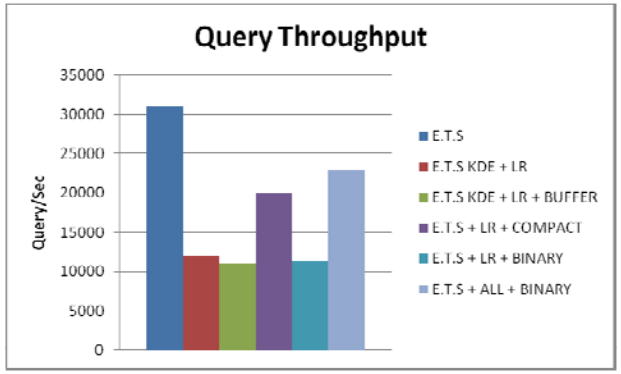


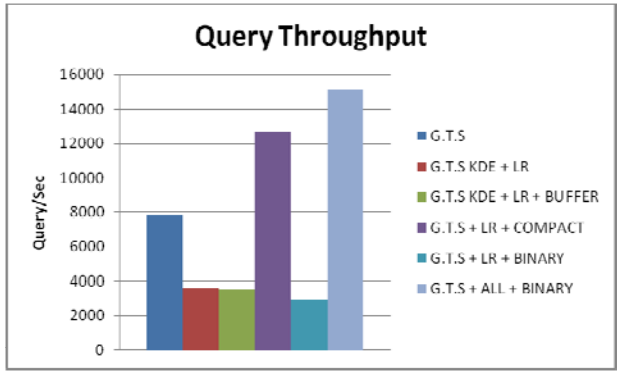


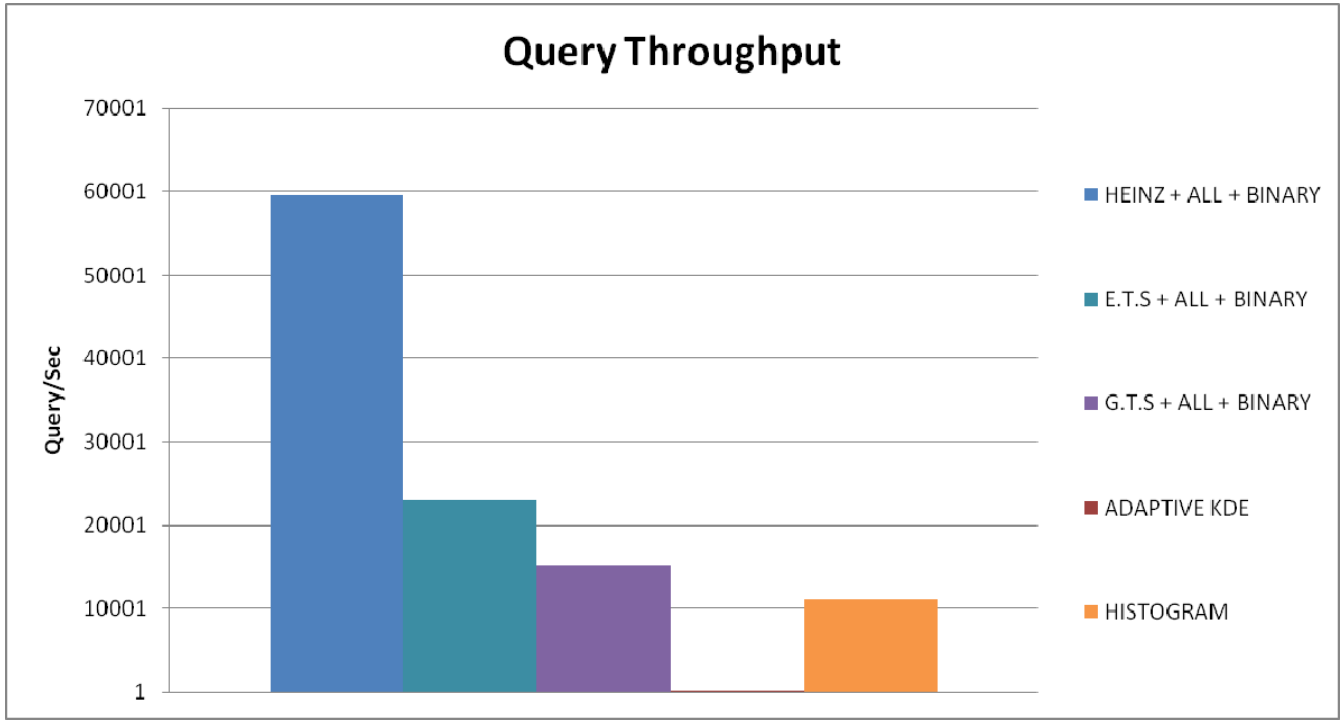


Query Throughput (MIX2-MIX8)

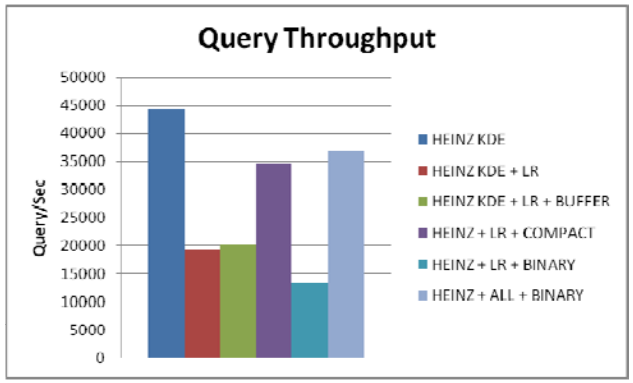


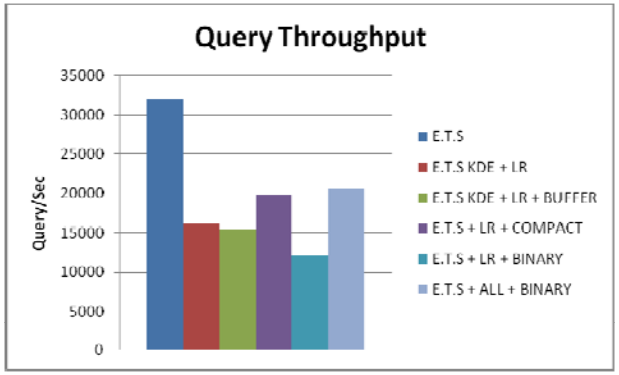


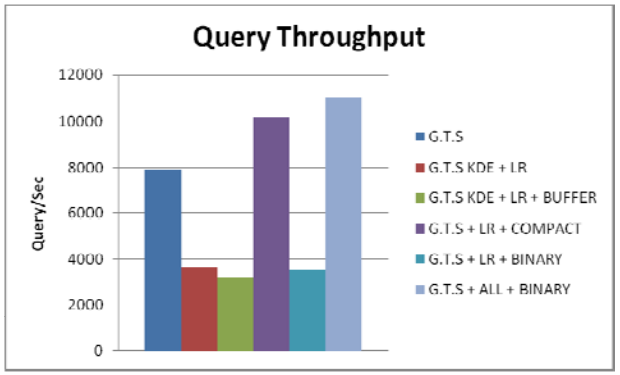


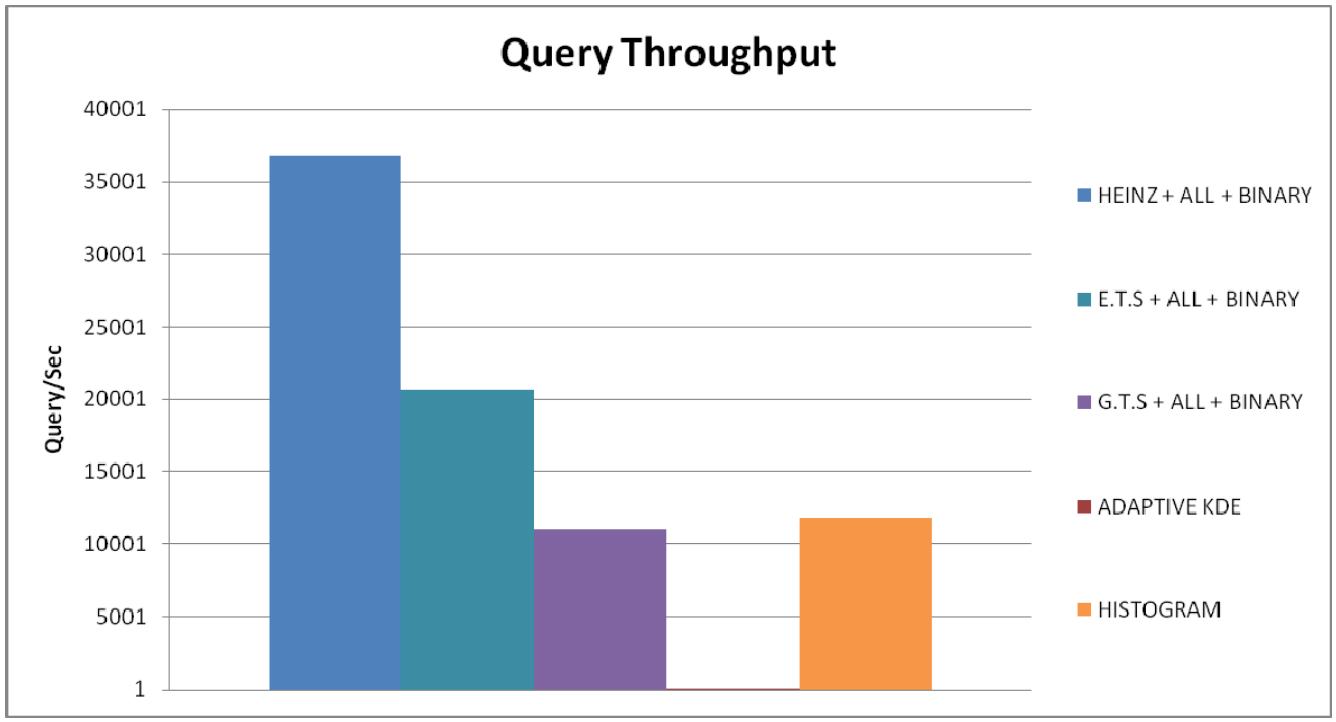


Query Throughput (POWER)

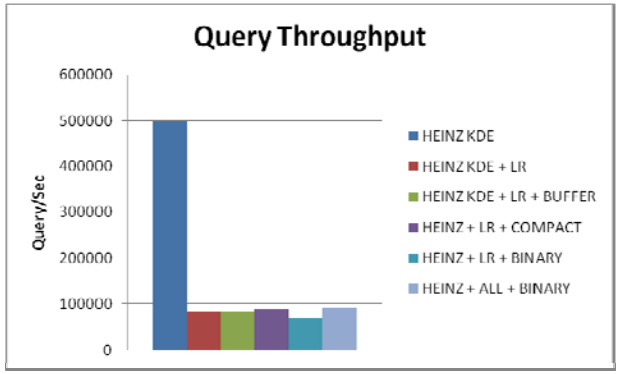


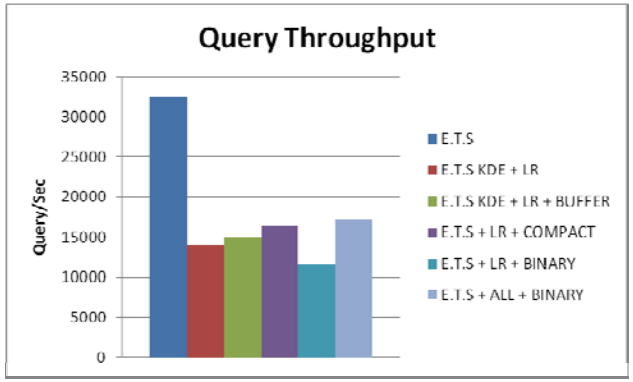


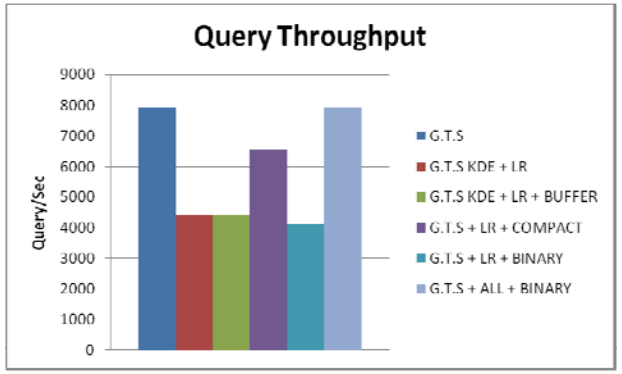


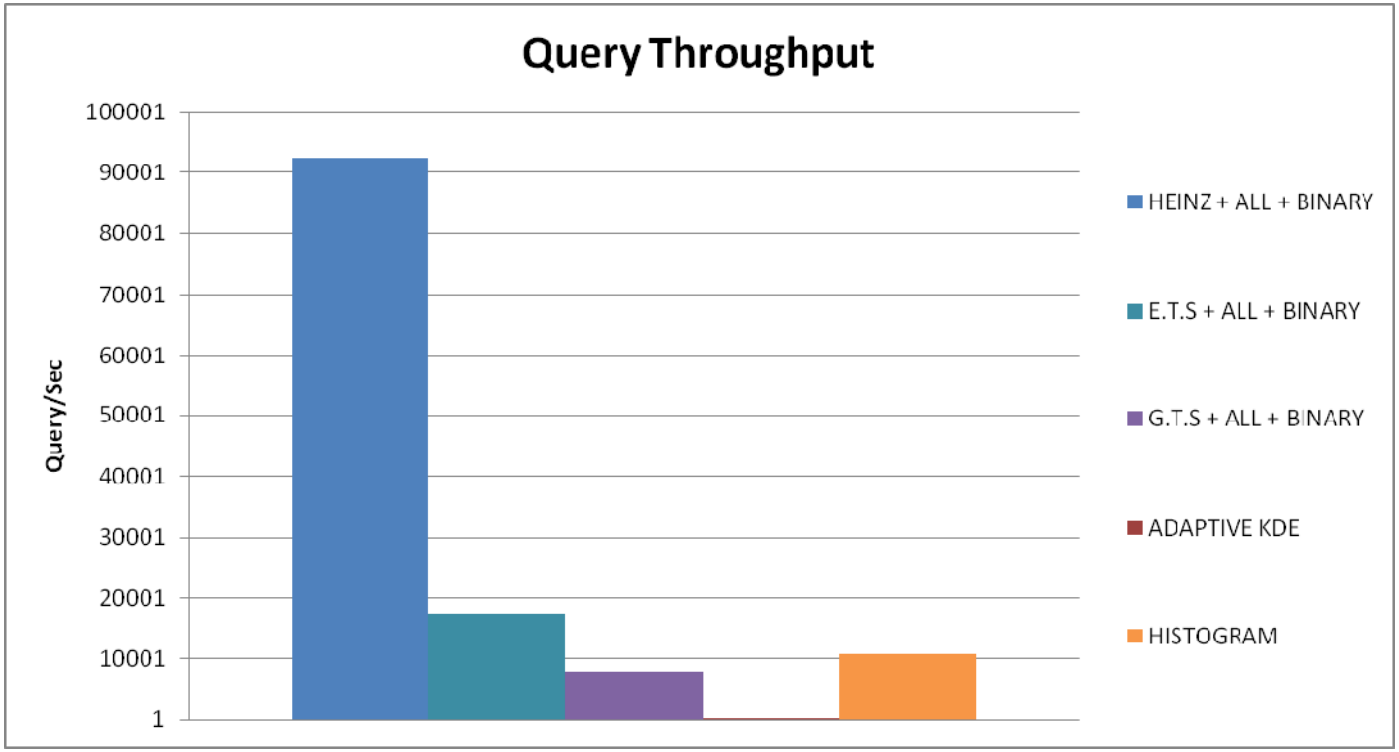


Query Throughput (TRAFFIC)









Schedule

- 7/1 – 7/8 – Draft 1
- 7/8 – 7/23 – Revision and possible experiment extension
- 7/24 – 7/31 – Refinement
- 8/1 – due date for Proceedings of VLDB