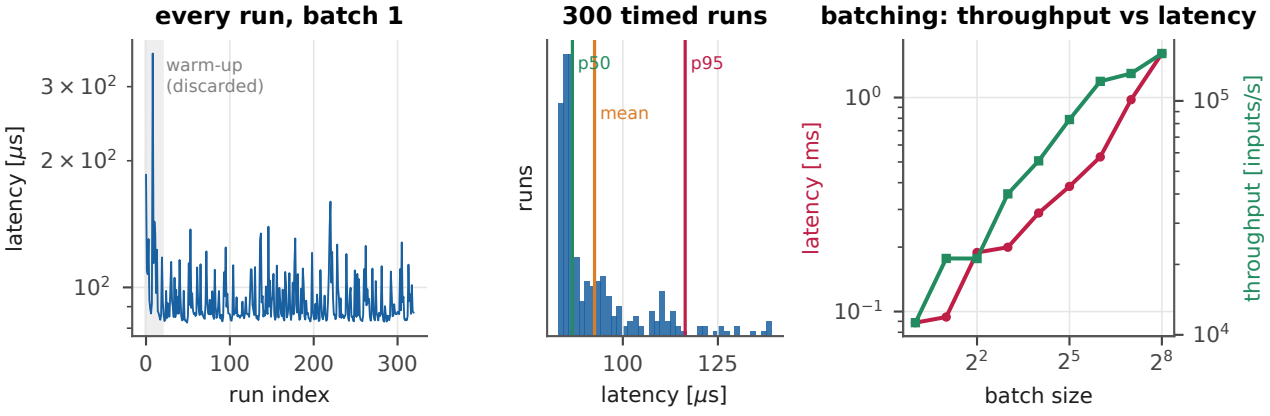

every run, batch 1
warm-up
(discarded)
3×10^2
2×10^2
10^2
latency [μs]
0
100
200
300
run index
300 timed runs
p50
p95
mean
runs
100
125
latency [μs]
batching: throughput vs latency
10^0
10^{-1}
latency [ms]
10^5
10^4
throughput [inputs/s]
2^2
2^5
2^8
batch size