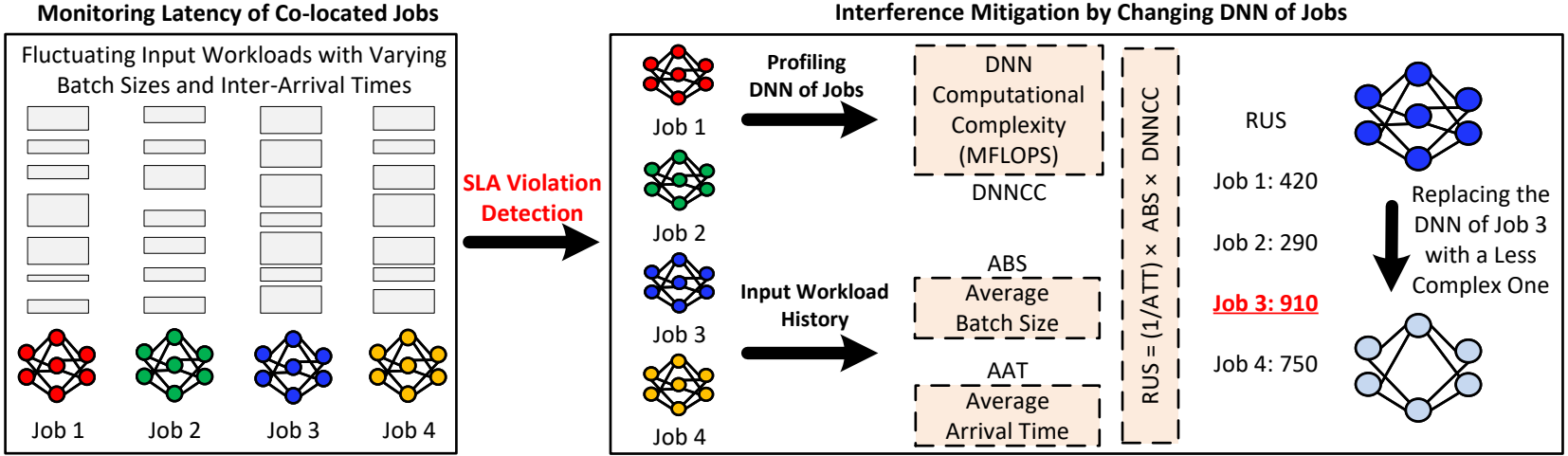
Monitoring Latency of Co-located Jobs
Fluctuating Input Workloads with Varying Batch Sizes and Inter-Arrival Times
Job 1
Job 2
Job 3
Job 4
SLA Violation Detection
Interference Mitigation by Changing DNN of Jobs
Job 1
Job 2
Job 3
Job 4
Profiling DNN of Jobs
Input Workload History
DNN Computational Complexity (MFLOPS)
DNNCC
ABS
Average Batch Size
AAT
Average Arrival Time
RUS = (1/ATT) × ABS × DNNCC
RUS
Job 1: 420
Job 2: 290
Job 3: 910
Job 4: 750
Replacing the DNN of Job 3 with a Less Complex One