Autoscaling

All my servers are completely busy right now. Consider adding more servers!

web service

add and remove servers manually

- doing it often is operationally expensive
- infrequent changes increase the likelihood of overload
- keeping excess capacity in the cluster increases dollar cost

add and remove servers automatically

- improves availability
- reduces costs
- improves performance (both latency and throughput)

Autoscaling

performance metrics are used
to make decisions on whether to scale or not

metric-based

scale out
when average CPU utilization > 80%

scale in
when average CPU utilization < 40%

- CPU utilization
- memory utilization
- disk utilization
- request count
- active threads count

we define a schedule for the
autoscaling system to follow

schedule-based

scale out
during business hours

scale in
at night and on weekends

machine learning models are used
to predict expected traffic

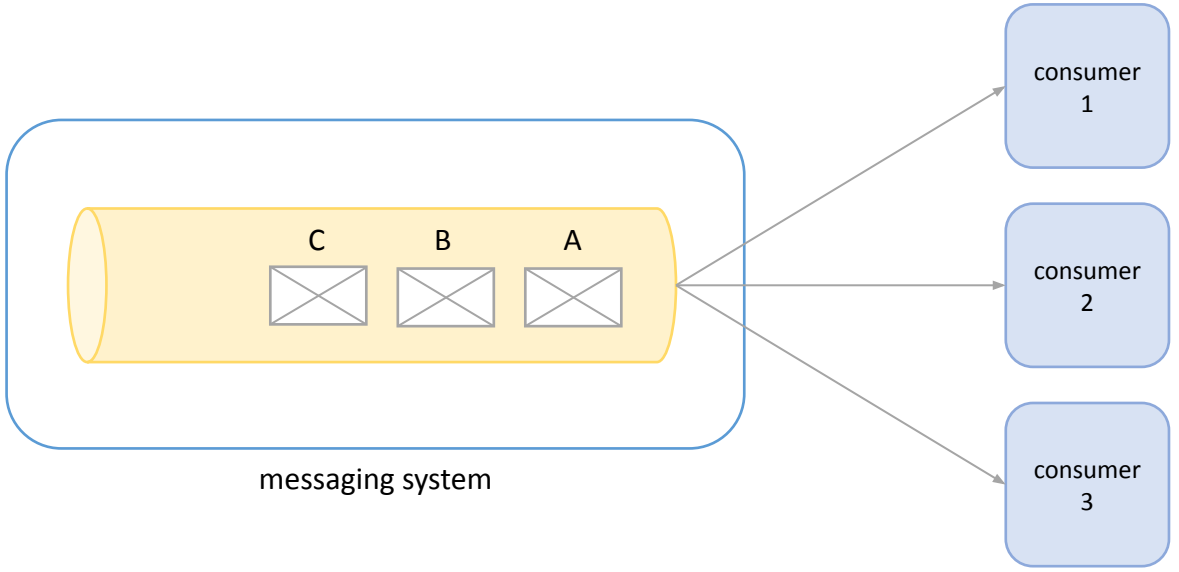
predictive

scale out
when average CPU utilization is expected to increase

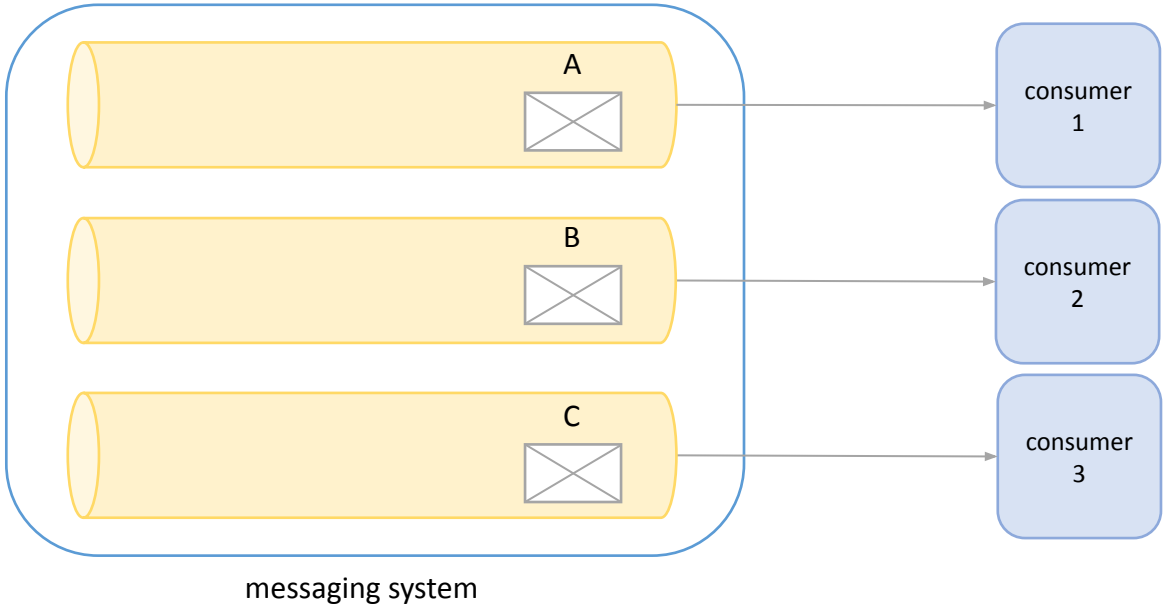
scale in
when average CPU utilization is expected to decrease

multiple scaling policies can be
used together

Autoscaling



add and remove consumer instances



add (split) and remove (merge) partitions