Scaling of EC2 Using SQS

Problem Statement:

In this scenario, you're a solutions architect at an e-commerce firm. The company runs flash sales from time to time, and when there's a spike in orders, the fulfillment backend can struggle to meet demand. One way to solve the problem is to overprovision EC2 instances in the fulfillment system to provide headroom to process all the orders. However, this can be very costly, since you'll have unused capacity when the traffic subsides.

Solution

What is the better way? Well, there is, create Auto Scaling rules for EC2 based on the number of messages in an SQS queue.