Appendix

1 Calculating number of VMs

In the Always On (AO) model, recall that we have two crossover points: L_s and L_v. If $L_v \leq L_s$, then we always use serverless. On the other hand, if $L_v > L_s$, the workload (λ) will be either sent to serverless if $\lambda \leq L_s$ or sent to VM if $L_s < \lambda \leq L_v$.

We now derive the equation for calculating number of VMs required for a workload (λ) in the latter case where $L_s < \lambda \leq L_v$.

$$\text{Maximum load on VM} = L_v \tag{1}$$

From (1), we can calculate the minimum number of VMs as follows.

$$\text{Minimum number of VMs} = \lfloor \lambda / L_v \rfloor \tag{2}$$

From (1) and (2), maximum workload possible on $\lfloor \lambda / L_v \rfloor$ VMs is shown below.

$$\text{Maximum load possible} = \lfloor \lambda / L_v \rfloor L_v \tag{3}$$

Now, we compute the excess load to be sent to either serverless or VM using (3) as follows:

$$\text{Excess load}(\epsilon) = \lambda - \lfloor \lambda / L_v \rfloor L_v \tag{4}$$

If $\epsilon \leq L_s$, we send the excess workload to serverless. Otherwise, in the case when $L_s < \epsilon \leq L_v$, we instantiate an additional VM for the excess workload. Therefore, using (2) and (4), we determine the required number of VMs as follows:

$$v(\lambda) = \begin{cases} \left\lfloor \frac{\lambda}{L_v} \right\rfloor, & \text{if } \lambda - \lfloor \lambda / L_v \rfloor L_v \leq L_s \\ 1 + \left\lfloor \frac{\lambda}{L_v} \right\rfloor, & \text{otherwise} \end{cases} \tag{5}$$

In the On-Off (OO) model, we similarly derive the equation for number of VMs. In this model, we have three crossover points: $L_s^{'}$, $L_v^{'}$ and L_v. If $L_v \leq L_s^{'}$, then we always use serverless. If $L_v^{'} \leq L_s^{'} < L_v$, then we switch directly from serverless to VM always on as the load increases. This case is identical to AO model and uses the equation 5 to calculate the number of VMs. If $L_s^{'} < L_v^{'} < L_v$, then we switch from serverless to VM on-off to VM always on with increase in load. However, as noted earlier in the model, serverless might not be used at all depending on the parameter values. For this final case, $L_s^{'} < L_v^{'} < L_v$, we now derive the equation for calculating number of VMs required for a workload (λ). The excess workload (ϵ) is calculated to be $\lambda - \lfloor \lambda / L_v \rfloor L_v$ where $\lfloor \lambda / L_v \rfloor$ still represents the minimum number of VMs required. There are three cases depending on the value of excess workload to determine the compute used. In the first case, $\epsilon \leq L_s^{'}$, we send the excess workload to serverless. In the second case, $L_s^{'} < \epsilon \leq L_v^{'}$, we send the excess workload to On-Off VM. Finally, if $L_v^{'} < \epsilon \leq L_v$, then we instantiate an additional always on VM.

2 Calculating serverless load

In the AO model where $L_v > L_s$, we only send the excess load to serverless when $\epsilon \leq L_s$. Otherwise, we send the entire excess load to a VM. Therefore, from (4), we can compute the load sent to serverless as follows:

$$s(\lambda) = \begin{cases} \lambda - \left\lfloor \frac{\lambda}{L_v} \right\rfloor L_v, & \text{if } \lambda - \lfloor \lambda / L_v \rfloor L_v \leq L_s \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

3 Expected cost of VM rental with On-Off dynamics (OO)

In the OO model, in a single cycle of start-up delay, a busy period, and an idle period, the cost of a VM which is turned on and off depends on the time for which the VM is on. This time includes busy period and the idle period of the VM. The expected time a.k.a. active time a VM is on is mentioned below.

$$\text{Active time} = E[startup]+E[busy] = 1/\gamma+1/(\mu_v-\lambda) \tag{7}$$

We can represent total time of a single cycle as follows.

$$\begin{aligned}\text{Total time} &= E[startup] + E[busy] + E[idle] \\ &= 1/\gamma + 1/(\mu_v - \lambda) + 1/\lambda\end{aligned} \tag{8}$$

Therefore, from (7) and (8), we compute the expected cost of VM rental as follows.

$$c(\lambda) = \alpha_v \frac{\text{Active time}}{\text{Total time}} = \alpha_v \frac{1/\gamma + 1/(\mu_v - \lambda)}{1/\gamma + 1/(\mu_v - \lambda) + 1/\lambda} \tag{9}$$

4 Extending busy period

We want to calculate the service capacity $\mu_v^{'}$ of a modified server under a load λ such that its expected busy period is given by $1/\gamma + 1/(\mu_v - \lambda)$.

The expected busy period of this modified server is obtained using the M/M/1 model as follows.

$$\text{E[busy]} = 1/(\mu_v^{'} - \lambda) \tag{10}$$

Therefore, from (10), we get

$1/(\mu_v^{'} - \lambda) = 1/\gamma + 1/(\mu_v - \lambda)$

$1/(\mu_v^{'} - \lambda) = \frac{\mu_v - \lambda + \gamma}{\gamma(\mu_v - \lambda)}$

$(\mu_v^{'} - \lambda) = \frac{\gamma(\mu_v - \lambda)}{\mu_v - \lambda + \gamma}$

$\mu_v^{'} = \lambda + \frac{\gamma(\mu_v - \lambda)}{\mu_v - \lambda + \gamma}$

We obtain the value of $\mu_v^{'}$ by simplifying the above equation as follows.

$$\mu_v^{'} = \frac{\lambda(\mu_v - \lambda) + \gamma\mu_v}{\mu_v - \lambda + \gamma} \tag{11}$$