APPENDIX A
DATA LOCALITY AWARE TASK SCHEDULER

A.1 Formula Function for Remaining Execution Time Estimation

ReLu activation function $f(x)$ is improved as:

$$f(x) = \begin{cases} \frac{x+|x|}{2}, \alpha = 0 \\ \frac{1+\alpha}{2}x + \frac{1-\alpha}{2}|x|, \alpha \neq 0 \end{cases} \quad (1)$$

where x is symbolic tensor to compute the activation function, α is scalar of tensor, which is optional, aiming to slope for negative input, usually between 0 and 1. The default value of 0 will lead to the standard rectifier, 1 will lead to a linear activation function, and any value in between will give a leaky rectifier.

APPENDIX B
DEADLINE CONSTRAINED JOB SCHEDULER

B.1 State Steady Equation

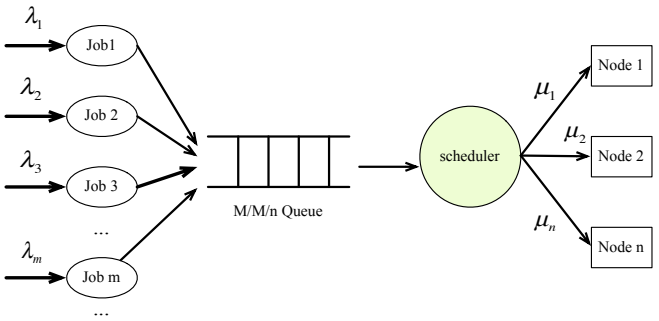

Fig. 1: M/M/n queuing model for job performing in cloud computing

Fig. 1 shows the M/M/n queuing model in cloud computing. Job requests come from different tenants and a cloud system should provide its services continually and will not restrict the number of the jobs. The state set of the cloud system is $\Phi = \{0, 1, 2, ...\}$, so these balance equations can be derived by the state transition flow diagram of the M/M/n queuing model depicted in Fig. 2. It shows the probability of different system's status and servers' status. When the state is k $(0 < k \leq n)$, k servers are busy and the remaining $n-k$ servers are idle. When the state is $k > n$, all the n servers are busy, and $k-n$ jobs are waiting for the service. Assume that there are 2 waiting queues: deadline queue and regular queue.

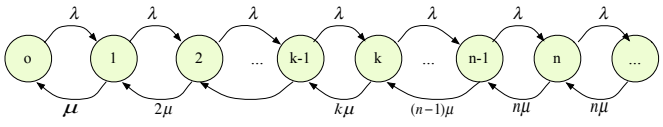

Fig. 2: State transition flow diagram of M/M/n queuing model

When the system is stable, let $\rho = \frac{\lambda}{n\mu}$ and the stability condition $\rho < 1$. The stationary probability p_k for state k can be determined by solving the set of balance equations, which state that the flux into a state should be equal to the flux out of this state when the system is stationary [1]:

$$\begin{aligned} &\text{When state } k = 0, \lambda p_0 = \mu p_1, \\ &\quad p_1 = \rho_1 p_0 = n\rho p_0; \\ &\text{When state } k = 1, \lambda p_1 = 2\mu p_2, \\ &\quad p_2 = \frac{\rho_1^2}{2!}p_0 = \frac{n^2}{2!}\rho^2 p_0; \\ &\text{When state } k = 2, \lambda p_2 = 3\mu p_3, \\ &\quad p_3 = \frac{\rho_1^3}{3!}p_0 = \frac{n^3}{3!}\rho^3 p_0; \\ &\quad \cdots \\ &\text{When state } k = n-1, \lambda p_{n-1} = n\mu p_n, \\ &\quad p_n = \frac{\rho_1^n}{n!}p_0 = \frac{n^n}{n!}\rho^n p_0; \\ &\text{When state } k = n, \lambda p_n = n\mu p_{n+1}, \\ &\quad p_{n+1} = \frac{\rho_1^{n+1}}{n!n}p_0 = \frac{n^n}{n!}\rho^{n+1} p_0; \\ &\quad \cdots \\ &\text{When state } k = n+r-1, \lambda p_{n+r-1} = n\mu p_{n+r}, \\ &\quad p_{n+r} = \frac{\rho_1^{n+r}}{n!n^r}p_0 = \frac{n^n}{n!}\rho^{n+r} p_0. \end{aligned} \quad (2)$$

In general,

$$p_k = \begin{cases} \frac{\rho_1^k}{k!}p_0 = \frac{n^k}{k!}\rho^k p_0, & 0 \leq k < n \\ \frac{\rho_1^k}{n!n^{k-n}}p_0 = \frac{n^n}{n!}\rho^k p_0; & k \geq n \end{cases} \quad (3)$$

According to regularity condition $\sum_{k=0}^{\infty} p_k = 1$, when $\rho < 1$, we can get

$$1 = (\sum_{k=0}^{n-1}\frac{\rho_1^k}{k!} + \sum_{k=n}^{\infty}\frac{\rho_1^k}{n!n^{k-n}})p_0 = (\sum_{k=0}^{n-1}\frac{\rho_1^k}{k!} + \frac{\rho_1^n}{n!}\frac{1}{1-\rho})p_0$$

Thus, we can gain an equation for p_0

$$p_0 = (\sum_{k=0}^{n-1}\frac{\rho_1^k}{k!} + \frac{\rho_1^n}{n!}\frac{1}{1-\rho})^{-1} \quad (4)$$

B.2 Density Functions and Distribution Functions of Sojourn Time and Waiting Time

The variance of mean number of jobs waiting in the queue shows as follows. Since,

$$\begin{aligned} &E(\bar{L}_{wai}^2) = \sum_{k=n}^{\infty}(k-n)^2 p_k = \sum_{h=1}^{\infty} h^2 p_{h+n} \\ &= \sum_{h=1}^{\infty}\frac{h^2}{n!n^h}(n\rho)^{h+n}p_0 \\ &= \frac{(n\rho)^n\rho^2 p_0}{n!}\sum_{h=2}^{\infty}h(h-1)\rho^{h-2} + \frac{(n\rho)^n\rho p_0}{n!}\sum_{h=1}^{\infty}h\rho^{h-1} \\ &= \frac{2\rho^2\rho_1^n p_0}{n!(1-\rho)^3} + \bar{L}_{wai} = \frac{1+\rho}{1-\rho}\bar{L}_{wai} \end{aligned} \quad (5)$$

Thus,

$$\sigma^2(\bar{L}_{wai}) = E(\bar{L}_{wai}^2) - [E(\bar{L}_{wai})]^2 = \bar{L}_{wai}(\frac{1+\rho}{1-\rho} - \bar{L}_{wai}) \quad (6)$$

In addition,

$$E(L_{sys}) = E(L_{wai}) + E(L_{ser}) \quad (7)$$

$$E(W_{soj}) = E(W_{wai}) + E(T_{ser}) \quad (8)$$

we can easily get

$$E(T_{ser}) = \frac{1}{\mu} \quad (9)$$

If we only consider servers of the system, without regarding the waiting queues outside the servers, it is easy to observe that there are no losses, and therefore the arrival rate in this cloud

system is λ, and the mean waiting time of each customer is $E(T_{ser}) = \frac{1}{\mu}$ [2], [3].

To obtain $E(L_{wai}|q \geq n)$, noted that the evolution of the M/M/n queue during the time when $q \geq n$ is equal to that of M/M/1 queue with the arrival rate λ and the service rate $n\mu$. Therefore, the mean queue length of this kind of M/M/1 queue is equivalent to $\frac{1}{1-\rho}$, where $\rho = \frac{\lambda}{n\mu}$. Therefore,

$$E(L_{wai}|q \geq n) = \frac{\rho/n}{1-\rho/n} = \frac{\rho}{n-\rho} \tag{10}$$

Due to $E(L_{wai}|q < n) = 0$, $P(q \geq n) = C(n, \rho)$, we get

$$E(L_{wai}) = C(n, \rho)\frac{\rho}{n-\rho} \tag{11}$$

The following formulas calculate the distribution of waiting time and sojourn time. An arriving job has to wait if at its arrival the number of jobs in the system is at least n and thus the time while a customer is serviced is exponentially distributed with parameter $n\mu$. Consequently if there are $n+j$ jobs in the system, the waiting time is Erlang distributed with parameters $(j+1, n\mu)$. By applying the theorem of total probability to the density function of waiting time we get [3]

$$f_w(x) = \sum_{j=0}^{\infty} p_{n+j}(n\mu)^{j+1}\frac{x^j}{j!}e^{-n\mu x} \tag{12}$$

Substitute the distribution for the density function of the waiting time, we get

$$\begin{aligned} f_w(x) &= \tfrac{p_0(\frac{\lambda}{\mu})^n}{n!} n\mu e^{-n\mu x} \sum_{j=0}^{\infty} \tfrac{(\rho n\mu x)^j}{j!} \\ &= \tfrac{(\frac{\lambda}{\mu})^n}{n!} p_0 n\mu e^{-(n\mu-\lambda)x} \\ &= \tfrac{(\frac{\lambda}{\mu})^n}{n!} p_0 n\mu e^{-n\mu(1-\rho)x} \\ &= \tfrac{(\frac{\lambda}{\mu})^n}{n!} p_0 \tfrac{1}{1-\rho} n\mu(1-\rho) e^{-n\mu(1-\rho)x} \\ &= P(Waiting) n\mu(1-\rho) e^{-n\mu(1-\rho)x} \end{aligned} \tag{13}$$

Thus for the complement of the the distribution function, we have

$$\begin{aligned} P(W > x) &= \int_x^{\infty} f_w(u)du = P(Waiting)e^{-n\mu(1-\rho)x} \\ &= C(n,\rho) \bullet e^{-\mu(n-\frac{\rho}{n})x} \end{aligned} \tag{14}$$

The distribution function of waiting time can be written as:

$$\begin{aligned} F_w(x) &= 1 - P(Waiting) + P(Waiting)(1 - e^{-n\mu(1-\rho)x}) \\ &= 1 - P(Waiting)e^{-n\mu(1-\rho)x} \\ &= 1 - C(n,\rho) \bullet e^{-\mu(n-\frac{\rho}{n})x} \end{aligned} \tag{15}$$

If the arriving number of jobs in the system is smaller than n, then the jobs will immediately get serviced. Otherwise, the jobs have to wait and their sojourn times include waiting time and service time. By applying the law of total probability to the density function of sojourn time, $f_s(x)$ is given as follows:

$$f_s(x) = P(No\ waiting)\mu e^{-\mu x} + f_{w+ser}(x) \tag{16}$$

Whereas, the density function of sojourn time for the job that needs to wait first $f_{w+ser}(x)$:

$$\begin{aligned} f_{w+ser}(z) &= \int_0^z f_w(x)\mu e^{-\mu(z-x)}dx \\ &= P(Waiting)n\mu(1-\rho)\mu \int_0^z e^{-n\mu(1-\rho)x}e^{-\mu(z-x)}dx \\ &= \tfrac{(n\rho)^n}{n!} p_0 \tfrac{1}{(1-\rho)} n\mu(1-\rho)\mu e^{-z\mu} \int_0^z e^{-\mu(n-1-\frac{\lambda}{\mu})x}dx \\ &= \tfrac{(n\rho)^n}{n!} p_0 n\mu \tfrac{1}{(n-1-\frac{\lambda}{\mu})} e^{-z\mu}(1 - e^{-\mu(n-1-\frac{\lambda}{\mu})z}) \end{aligned} \tag{17}$$

For the complement of the distribution function of the response time, we get

$$\begin{aligned} &P(S > x) = \int_x^{\infty} f_s(y)dy = \\ &\int_x^{\infty} \mu e^{-\mu y} + \tfrac{(\frac{\lambda}{\mu})^n p_0}{n!(1-\rho)} \tfrac{1}{(n-1-\frac{\lambda}{\mu})}(\mu e^{-\mu y} - \mu(n-\tfrac{\lambda}{\mu})e^{-\mu(n-\frac{\lambda}{\mu})y})dy \\ &= e^{-\mu x} + (\tfrac{\lambda}{\mu})^n p_0 \tfrac{1}{n!(1-\rho)(n-1-\frac{\lambda}{\mu})}(e^{-\mu x} - e^{-\mu(n-\frac{\lambda}{\mu})x}) \\ &= e^{-\mu x}(1 + \tfrac{(\frac{\lambda}{\mu})^n p_0}{n!(1-\rho)} \tfrac{1-e^{-\mu(n-1-\frac{\lambda}{\mu})x}}{(n-1-\frac{\lambda}{\mu})}) \end{aligned} \tag{19}$$

Therefore the distribution function can be presented as

$$F_s(x) = 1 - P(S > x) \tag{20}$$

REFERENCES

[1] W. Ellens, J. Akkerboom, R. Litjens, and H. Van Den Berg, "Performance of cloud computing centers with multiple priority classes," in *Proceedings of 5th IEEE International Conference on Cloud Computing (CLOUD)*. IEEE, 2012, pp. 245–252.

[2] L. Guo, T. Yan, S. Zhao, and C. Jiang, "Dynamic performance optimization for cloud computing using m/m/m queueing system," *Journal of Applied Mathematics*, vol. 2014, 2014.

[3] J. Sztrik, "Basic queueing theory," *University of Debrecen, Faculty of Informatics*, vol. 193, 2012.