Introduction to Queueing Theory

Prof. N. Selvaraju
Department of Mathematics
Indian Institute of Technology Guwahati, India

Lecture - 23
Erlangian Queues: Erlangian Arrivals, Erlangian Service Times

Hi and hello, everyone. Having seen the distributions that are more general than exponential, that is the Erlang hyper-exponential hypo-exponential or, in general, a general phase-type distribution. Now we will look at a couple of models utilizing these kinds of distribution. So, mainly we will be looking at only the Erlangian model. So, basically, we will look at the Erlang-based queueing models. And how one can analyze that queueing model is what then we will see in this lecture. Now, the idea that we gain from here is, in general, applicable to general phase-type distribution, and you can see how we are getting back to the Markovian nature in some form; everything will be explicit in this discussion. But since we are not going to see a model which is based upon a general phase-type distribution, the same idea is what we can be extended to general phase-type distribution. And hence like, you pay attention to like how we are dealing with this and how we are constructing the Markov chain in this particular case. Now, you see, what we start with is a model which we call the Erlang Service Model, meaning that we have an $M/E_k/1$ queue. So, what do we have?

- All the assumptions of an $M/M/1$ model (in equilibrium) are in place, except that the service time distribution is now an Erlang type-k (E_k) distribution (with mean $\frac{1}{\mu}$).

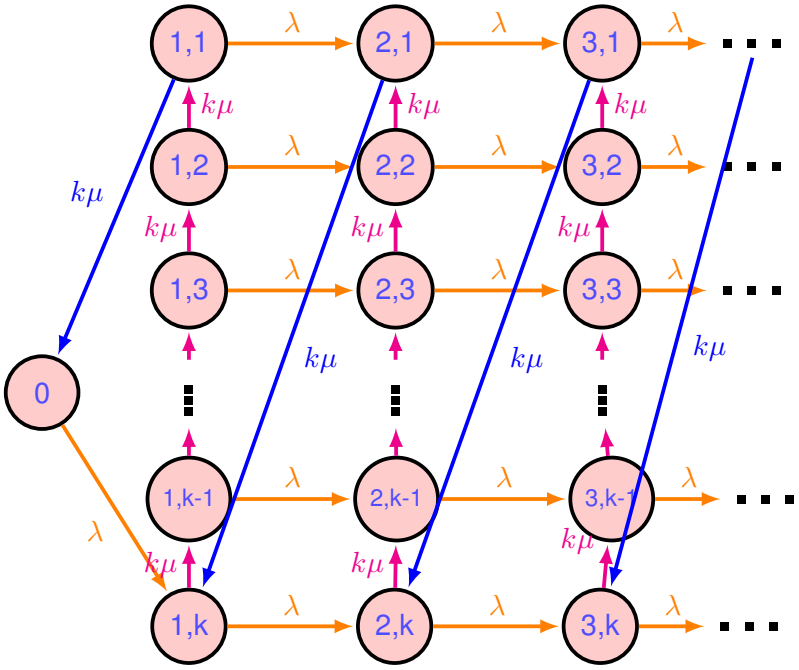
  We retain this mean $1/\mu$ for ease. All the assumptions of the $M/M/1$ model in equilibrium are in place except that the service time distribution now is replaced with an Erlang type-k or E_k distribution with mean $1/\mu$. We retain this mean $1/\mu$ for ease. So, that is how then we have defined this Erlang type-k distribution as well, and this is what is the use or the utility that comes in handy here. So, this is the difference; all assumptions of $M/M/1$ are in place except that the service time distribution now is replaced by an Erlang distribution rather than an exponential distribution, but since Erlang, you can view it as the sum of the i.i.d exponential.

- The service time duration may be thought of as consisting of k independent and identical exponential phases (or stages), each with mean $\frac{1}{k\mu}$ (even if there are not phases, it is convenient to think this way).
- Since the overall service rate is assumed to be μ, the rate of each service phase is $k\mu$.
- The system state can be described by a two-dimensional CTMC with state space $S = \{0\} \cup \{(n, i) : n \geq 1, 1 \leq i \leq k\}$, where n denotes the number of customers in the system and i is the current phase of the service for the customer currently in service.
- The phases are numbered backwards, so k is the first phase of service and 1 is the last phase.

So, essentially like when you are looking at here, you see that this total is Erlang, but we call this has the kth phase and so on, and this has the 1st phase and the service time. So, this total duration is what is my E_k, that is what you have here, and each of these has a rate k. So, this is what we are looking at it. So, the first phase is k is what we see, and 1. Now, suppose if it is in say some ith phase, meaning that i more phases to go in this service, that is what it would mean. So, when it is in 1 at the end of this phase, his service is getting completed; that is how we have to view it. So, it is important to understand that in order to describe this Markov chain and how it behaves and so on. So, that is the part for that convenience; just anytime if it is in phase i, means i more phases are remaining. It is no different than if you order it in the usual way, then it will be $k - i$ is what $i + 1$ you have to write. So, it is easier to write it i, so that is what is given.

- The total number of phases left when the system is in state (n, i) equals $(n - 1)k + i$.

So, if these phases are ordered, number differently, then this will be changing; that is a difference, so that is what we have here. So, this is the Markov chain 2-dimensional Markov chain; we can call it one as the number and the other dimension as the phase of the current service which is undergoing. So, that is what is this (n, i); remember this n and i. Now, let us look at how this system would behave, and let us try to draw the state transition diagram of how the system will behave in this case.



- We can write the following steady-state balance equations

$$
\begin{aligned}
(\lambda + k\mu)p_{n,i} &= k\mu p_{n,i+1} + \lambda p_{n-1,i}, \quad n \geq 2, \quad 1 \leq i \leq k-1 \\
(\lambda + k\mu)p_{n,k} &= k\mu p_{n+1,1} + \lambda p_{n-1,k}, \quad n \geq 2 \\
(\lambda + k\mu)p_{1,i} &= k\mu p_{1,i+1}, \quad 1 \leq i \leq k-1 \\
(\lambda + k\mu)p_{1,k} &= k\mu p_{2,1} + \lambda p_0, \\
\lambda p_0 &= k\mu p_{1,1}
\end{aligned}
$$

You have written down the flow balance equations, if this is a reasonable system, obviously, then you will be able to handle it, and you can get the solution, but at least you have written down; now, you can adopt an appropriate method to do this.

- These equations are not that easy to handle due to their bivariate nature.
- Interestingly, one can relate this Erlangian service queue to a constant bulk input model $M^{[K]}/M/1$, where each input unit brings in $K = k$ phases and the (phase) service rate μ is replaced by $k\mu$.

  So, you can make an equivalence as if it brings in each arrival because if you look at the number of phases, we said that you go back, we said that number of phases here. So now, if this is the state; now each arrival will bring in k, and if I look at $(n-1)k+i$, that is where this transformation if I make the transformation, it might be useful for which only this relationship is, or this equivalence between these two systems is important is that each arrival you can think as if it brings in k phases, which in bulk, you can think if as if it is basically it brings in k arrivals each of them now the rate is $k\mu$, not μ that is you have to remember that. So, that rate is $k\mu$, not μ, because each of these phases requires $k\mu$ as the rate.
- One can utilize this equivalence of $M/E_k/1$ and $M^{[k]}/M/1$ to get the performance measures easily.
- To determine the average wait in queue W_q for the $M/E_k/1$ queue:
  ▶ First observe that the average number of phases in the $M/E_k/1$ queue is the same as the average number of customers in the analogous $M^{[X]}/M/1$ queue (which was equal to $\frac{k+1}{2}\frac{\rho}{1-\rho}$).
- Here, in the Erlangian model, the service $k\mu$ replaces μ in the bulk arrival model and hence $\rho = \dfrac{k\lambda}{k\mu} = \dfrac{\lambda}{\mu}$.
  ▶ Since the average time to process each phase is $\dfrac{1}{k\mu}$, the average wait in queue is the average number of phases in the system multiplied by $\dfrac{1}{k\mu}$. We get

$$W_q = \frac{1+\frac{1}{k}}{2}\frac{\rho}{\mu(1-\rho)}, \quad \rho = \frac{\lambda}{\mu}.$$

  For this model, rho is λ/μ, that is a utilization because your arrival rate is λ, and the service rate is μ though it is distributed as Erlang. So, the ρ is λ/μ is the utilization factor or traffic intensity for this particular system. So, with that change, that is a change that you have to adapt to. So, when you adapt that, then your W_q would become $\dfrac{1+\frac{1}{k}}{2}\dfrac{\rho}{\mu(1-\rho)}$.

  So basically, we considered this $M^{[X]}//M/1$ model, and we also considered the special case where X is the constant batch size, and in that case, we arrived at $\dfrac{k+1}{2}\dfrac{\rho}{1-\rho}$, and we said that $\dfrac{\rho}{1-\rho}$ is the quantity corresponding to $M/M/1$ and it is multiplied by $\dfrac{k+1}{2}$ factor. Now, what we have to make adjustments because now in the ρ, we have to make adjustments where μ is replacing by $k\mu$ which will bring you $\rho = \dfrac{k\lambda}{k\mu} = \dfrac{\lambda}{\mu}$ as your new ρ, and with that new ρ, $W_q = \dfrac{1+\frac{1}{k}}{2}\dfrac{\rho}{\mu(1-\rho)}$ is your wait in the queue which is the average waiting time in the queue.
- It follows in the usual manner that

$$L_q = \lambda W_q = \frac{1+\frac{1}{k}}{2}\frac{\rho^2}{1-\rho}, \quad L = L_q + \rho, \quad W = \frac{L}{\lambda} = W_q + \frac{1}{\mu}, \quad \rho = \frac{\lambda}{\mu}.$$

So, that is a change that you have to remember. So, you see how we are making the equivalence, we are making one system that is equivalent to another system, and we are getting done with our work in this case. Getting the performance measures, of course, other performance measures also you can obtain if you could always relate these two systems with this change in mind.

- We will now try to get the steady-state probabilities for which we use the fact that state (n, i) can be transformed equivalent to $(n-1)k+i$ in a single-variable system.
- Mapping (n, i) to $(n-1)k+i$, this represents the number of phases left in this service.

  But one thing you have to remember, in between when you are making the equivalence between this $M/E_k/1$ and this bulk arrival; is the service completions here in the $M/E_k/1$ model does not correspond to customer departures. That is what you have to remember it. Keep that in mind. Whereas in a bulk queue, each customer departs, but here the customer is not getting departed; that is, you have to remember that.

- The steady-state balance equations can be rewritten as

$$0 = -(\lambda + k\mu)p_{(n-1)k+i} + k\mu p_{(n-1)k+i+1} + \lambda p_{(n-2)k+i}, \qquad n \geq 1, \quad 1 \leq i \leq k$$
$$0 = -\lambda p_0 + k\mu p_1$$

  where any p with a negative subscript is assumed to be zero.

- The above system can be rewritten again in a simplified manner (starting from $n = 1, i = 1$ and proceeding sequentially) as

$$0 = -(\lambda + k\mu)p_n + k\mu p_{n+1} + \lambda p_{n-k}, \qquad n \geq 1$$
$$0 = -\lambda p_0 + k\mu p_1$$

  which is the same set of equations as for the bulk arrival queue for a constant batch size k and service rate $k\mu$.

- Defining $P(z) = \sum_{n=0}^{\infty} p_n z^n$ and proceeding as earlier, we get

$$P(z) = \frac{k\mu p_0(1-z)}{k\mu - (\lambda + k\mu)z + \lambda z^{k+1}}.$$

- Using $P(1) = 1$ to get $p_0 = \frac{\mu-\lambda}{\mu} = 1 - \rho, \quad \rho = \frac{\lambda}{\mu}$, we obtain

$$P(z) = \frac{k\mu(1-\rho)(1-z)}{k\mu - (\lambda + k\mu)z + \lambda z^{k+1}}$$

  from which we can obtain $\{p_n\}$ by expanding $P(z)$ as a power series in z (through a partial fraction expansion).

- Now, the p_n's obtained above gives the probability of the number of phases in the system. If $p_n^{(c)}$ denotes the probability that the number of customers in the system is n, then

$$p_n^{(c)} = \sum_{m=(n-1)k+1}^{nk} p_m, \qquad n = 1, 2, \ldots.$$

- Special case: $k = 1$. ($M/M/1$ queue)
  We have $P(z) = \dfrac{\mu(1-\rho)(1-z)}{\mu - (\lambda+\mu)z + \lambda z^2} = \dfrac{1-\rho}{1-\rho z}$ so that $p_n = (1-\rho)\rho^n, n \geq 0$.

  But here, p_n and $p_n^{(c)}$, the number in the customer, p_n is number in the phase is 1 because the number of phases is 1. So, $p_n = p_n^{(c)}$ as well. So, this is true in this case. Now, let us look at a couple of examples.

  Example. *[$M/E_4/1$]*

  - Suppose that there is a single-server service system to which the customers arrive according to a Poisson process with a mean rate of $16/h$.
  - The mean service time is 2.5 minutes with a standard deviation of $\frac{5}{4}$ minutes. It is thought that the Erlang distribution would be a reasonable assumption for the service time distribution.
  - How long a customer on an average must wait until getting into the service and how many customers are waiting for service?

  - The appropriate model is an $M/E_k/1$ model.

    $$\lambda = \frac{16}{60}, \quad \frac{1}{\mu} = 2.5, \quad \sigma^2 = \frac{1}{k\mu^2} = \frac{25}{16} \quad \Longrightarrow \quad k = 4.$$

    We therefore have an $M/E_4/1$ model with $\rho = \frac{2}{3}$ for which

    $$L_q = \frac{5}{8}\frac{\frac{4}{9}}{1-\frac{2}{3}} = \frac{5}{6}, \qquad W_q = \frac{60}{16}\frac{5}{6} = \frac{25}{8} \text{ minutes}.$$

Now, let us look at another example.

Example. *[$M/M/2$ Vs. $M/D/1$]*

- Components from a production line system that fail the quality control test arrive at a repair facility according to a Poisson process with a rate 18 per hour.
- The repair facility has two specialists and each can repair the component in an average of 5 minutes, with the repair time being exponentially distributed.
- The company is proposed with an alternative option of leasing one machine that can repair the components in exactly $2\frac{2}{3}$ minutes (i.e., no variation in repair time).
- Assume that the machine leasing cost is roughly equal to the salary and other benefit costs of the two staffs. Should the company lease the machine?

- We need to compare W and L under the two alternatives: $M/M/2$ Vs. $M/D/1$

– For $M/M/2$, $\lambda = 18/h, \mu = 12/h$ and this means that $W_q = \frac{3}{28}h = 6.4$ minutes, $W = 6.4 + 5 = 11.4$ minutes and $L = \lambda W = (18/60)(11.4) = 3.42$.

– For $M/D/1$, we use $\lim_{k\to\infty} M/E_k/1$. Given $\lambda = 18/h, \mu = \frac{3}{8}/min = 22.5/h$. Therefore, $W_q = \lim_{k\to\infty}\left(\frac{1+1/k}{2}\frac{\rho}{\mu(1-\rho)}\right) = \frac{\rho}{2\mu(1-\rho)} = \frac{4}{45}h = \frac{16}{3}$ *minutes,* $W = \frac{16}{3}+\frac{8}{3} = 8$ *minutes* *and* $L = \lambda W = (18/60)(8) = 2.4$.

– Thus, leasing the machine is preferable.

So, this is about the $M/E_k/1$ model; the next model that we will consider is basically the Erlang arrival model or $E_k/M/1$ model.

- We now consider a model where the interarrival times follow an Erlang type-k distribution with mean $\frac{1}{\lambda}$ (the rest being as usual).

- Like the previous model, here too an equivalence exists between $E_k/M/1$ and $M/M^{[k]}/1$.

- We can think as: An arrival is passing through k phases (each with mean $\frac{1}{k\lambda}$) before actually entering the system (again, a convenient device for analysis).
  ▶ The phases are numbered now frontward from 0 to $k-1$.

- Let $p_n^{(P)}$ denote the number of *arrival phases* in the system in steady state.
  ▶ The number of arrival phases in the system includes k phases for each customer who has already arrived (but not yet departed) as well as the completed phases corresponding to the next arrival (even though this customer has not officially "arrived").

- Then, the probability $p_n^{(c)}$ (the number of customers being n in steady state) is given by

$$p_n^{(c)} = \sum_{j=nk}^{nk+k-1} p_j^{(P)}$$

- This system is identical in structure to the full-batch bulk-service model. The rate balance equations for $p_n^{(P)}$ in this model are identical to the rate balance equations for p_n in the bulk-service model, except with λ is now replaced by $k\lambda$ (and K by k). Therefore, we get

$$p_j^{(P)} = \rho(1-r_0)r_0^{j-k}, \qquad j \geq k-1, \quad \rho = \frac{\lambda}{\mu},$$

where r_0 is the single root in $(0,1)$ of the characteristic equation

$$\mu r^{k+1} - (k\lambda + \mu)r + k\lambda = 0.$$

- Thus, we obtain, for $n \geq 1$,

$$\begin{aligned} p_n^{(c)} = \sum_{j=nk}^{nk+k-1} p_j^{(P)} &= \rho(1-r_0)(r_0^{nk-k} + r_0^{nk-k+1} + \cdots + r_0^{nk-1}) \\ &= \rho(1-r_0)r_0^{nk-k}(1 + r_0 + \cdots + r_0^{k-1}) \\ &= \rho(1-r_0^k)(r_0^k)^{n-1}. \end{aligned}$$

And, $p_0 = 1 - \rho$.

- Note that $p_n^{(c)}$ has a geometric form (as with the $M/M/1$), but with r_0^k as the geometric parameter instead of ρ.
- The performance measures can be obtained as usual as

$$L = \rho(1 - r_0^k) \sum_{n=1}^{\infty} n(r_0^k)^{n-1} = \rho(1 - r_0^k) \frac{1}{(1 - r_0^k)^2} = \frac{\rho}{1 - r_0^k}.$$

$$L_q = L - \rho, \quad W = \frac{L}{\lambda}, \quad \text{and} \quad W_q = W - \frac{1}{\mu}.$$

So, this is the usual way performance measures you can obtain; you can even obtain waiting time distribution which will be much similar to the $M/M/1$ case and so on. But, we will see will not go into that; rather, we will see an example.

Example. *[$E_2/M/1$]*

- Arrivals occur to a single-server queueing system with an E_2 distributed interarrival times with a mean interarrival time of 30 minutes.
- The service times are exponentially distributed with a mean of 25 minutes.
- Determine the steady state system size probabilities and expected value measures of effectiveness of the system.

- *Given $\lambda = 2/h, \mu = \frac{12}{5}/h$ and $k = 2$, the characteristic equation is $\frac{12}{5}r^3 - \frac{32}{5}r + 4 = \frac{4}{5}(3r^3 - 8r + 5) = (r - 1)(3r^2 + 3r - 5) = 0$ and it has a positive root in $(0, 1)$ given by $r_0 = (-3 + \sqrt{69})/6 = 0.8844$.*

  We then have

$$p_n^{(c)} = \rho(1 - r_0^k)(r_0^k)^{n-1} = (5/6)(0.2178)(0.7822)^{n-1} = (0.2320)(0.7822)^n, n \geq 1.$$

  and $p_0 = 1 - \rho = 1/6$.

  The mean system size is $L = \frac{\rho}{1 - r_0^k} = \frac{5/6}{1 - 0.7822} = 3.8261$ and $W = L/\lambda = 1.9131$ h.
  Similarly, $W_q = 1.9131 - 5/12 = 1.4964$ h and $L_q = 3.8261 - 5/6 = 2.9928$.

So, this is about the $E_2/M/1$ model; again, what we are seeing is these models can be analyzed using Markovian theory, meaning that you can write down the rate balance equations and you can solve them. But while doing that, you observe this $M/E_k/1$; you do not need to solve it separately. You can make an equivalence between the corresponding bulk service bulk arrival models whenever $E_j/M/1$ or $M/E_k/1$ models are considered. And you use that equivalence to directly get the final solution here; that is what we have done in this particular case. Even if you have $E_j/E_k/1$ model, suppose both of them are Erlangian. Now, two-dimension is not sufficient. Now, you need to go beyond two, three-dimension you have to keep, and then you have to work. So, the thing is that the model gets complex, but one can do an analysis. In general, a phase-type distribution like this idea can be utilized. Once you have this, then you can write down the Markov chain, and then you can write down the rate balance, and you can then onwards see how

you can solve this. So, a sophisticated method handling the matrix form of this is what would be required. So, that is what would is done when you are handling phase-type distribution. So, all these things you can represent in terms of phase-type distributions. So, you take a generic representation of phase-type in terms of alpha and q tilde, as we have described earlier. So, in that particular case using that using the matrix theory like you can like represent things in the matrix form, and you can deal with matrix form in a much more compact way. It is the only compact way; again, if you expand, it will be messy, but in a compact way, you can handle it. And that whole lot of ideas use what is called matrix geometric analysis or matrix analytic methods and so on. The starting point will be the taking that distributions of interarrival time or service times as phase-type distribution. Then you directly employ those kinds of methods; that is what is probably the current standard in the literature that you would find when you are dealing with more complex models.

But they are still within the Markovian framework; that is what you need to remember. So, this is there is an exhibition of how one can utilize basically non-Markovian distribution; you represent it in terms of exponential. So, that you can still utilize within the framework of Markovian models and analyze, you can analyze the models in the usual way as you do for a Markovian model. That is what we illustrated with these two Erlangian models. We may not consider anymore of these kinds of models, but we might generally, later on, we will deal with the general distribution itself; how one can do, we can do that. So, with this, we end this discussion of the Erlangian queues at this level.

Thank you, bye.