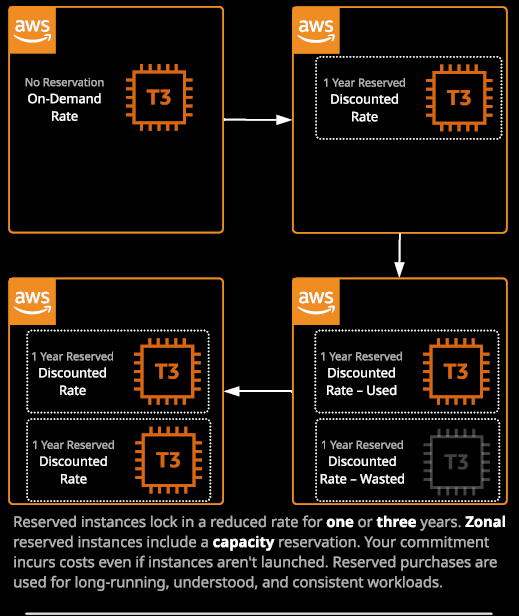
BILLING MODEL PART 2 – RESERVED INSTANCES  
  
So spot pricing was all about utilizing spare AWS capacity in order to get cheaper EC2 rates.

With reserved instances, it's kind of the inverse. What you doing **with the reserved purchases is you're giving AWS information about your future usage requirements.** You essentially telling AWS that you'll have a requirement to use a certain type of instance in a certain region or availability zone for a certain period of time and in exchange for that, you're getting these instances for cheaper. Now, the way the instances work is you purchase an instance reservation for a certain type of instance let's say it's a T3 and a certain size let's say that it's a T3.medium. Optionally you can also specify a specific availability zone to purchase that reservation in, but you don't have to. Reservations can be zonal or regional.



Now, the way that reserved instances work is that you can pay for reserved instances for a period of one year or three years. **You're essentially pre-purchasing the capacity for a given type and size of instance in a region or AZ for one or three years.** Purchasing it for three years means you commit to pay for that instance for three years. One year means you're committing to one year. So you purchase reserved instances by going to be EC2 console and then, under instances, reserved instances, you purchase reserved instances, so you'll need to select a particular platform type. Let's say Linux or Unix a particular type of instance. Maybe in this particular case, you want to go for an M5.large so I'll locate M5.large select that one. You need to pick a term to use we'll go for 12 to 36 months, and then you also need to select payment options. Now, remember, I just said that **you're purchasing an instance reservation for one or three years in either case, you committing to buy that capacity, but you've got an option about how you want to pay for it. You can either go for all upfront where you're paying for the entire term in one bulk payment that will offer you the maximum cost saving. You can go for partial upfront where you're paying a part of the fee upfront in exchange for a reduced hourly instant rate and finally can pay for no upfront, which just gives you a discount to the on demand pricing but you're still committed to pay it.**

**The key thing here is you can't switch off the instance stopping the instance or terminate the instance and not pay for this reservation. You pay for a reservation regardless.**

So this is an AWS account, and I'm just going to step through each of **four different scenarios**. So with the first box, this is just an **on demand pricing** for a T3 instance. So I've not bought any instance reservation. So I'm just paying the on demand rate. Now let's move on to this next box, where I've decided to buy a one year reserved purchase for this T3 instance. If I decided to pay all upfront, then I pay the entire year's costs for this instance and when I did that, I wouldn't have any ongoing costs there would be no hourly rate for this instance. I could pay no upfront and just have this for a discounted rate, but I'd still be committed to that one year term or I could pay a partial payment and then have a massively reduced hourly rate for this instance but the effect is the same. I'm getting a discount for running this T3 instance by committing to that one year in advance and paying for this instance whether it's in a started state a stop state or a terminated state it's the instance reservation that I'm paying for at this point rather than the instance. Now the way this works on a bill is that if you start up an instance that matches the type and size that the reservation is for, then you automatically either get the discounted rate or you don't pay an hourly rate, essentially the reservation think of that like a credit, and if you start matching instance, then you obviously you get a discount or you don't pay. Now let's move on to this next box. Imagine, in this case that I'd also purchased another one year reservation for the same type of instance. In that case, I'd be paying for this reservation, but I wouldn't be using it. It would be wasted. Now let's move on. If at that point, I started up another T3 instance which matched this wasted reservation that at that point I'd utilize this reservation, I need to get the discount or I'd be paying nothing if I paid for it all upfront. **Now the best way to reduce your bill if you can predict your usage over long term, is to purchase three year reservations and pay in its entirety upfront. In doing so, you get the best discount, and you don't have to worry about any hourly charges for the instances.**

Now, as well as deciding between the one year or three year and then the all upfront, partial upfront, or no upfront, you also get to decide whether you buy **a zonal reserved instance so that applies to a particular availability zone or a regional reservation**. If you choose to buy a zonal reserved instance, then you only benefit from that reservation in a particular availability zone but you also reserve capacity in that zone. Now more on this for a second, because capacity reservation is important. A zonal reserved instance that was also locked to a specific size.

So if you purchase an instance reservation, for example, as an M5.large. It'll only apply if you start up an M5.large. So that's key a zonal reserved instance reserves capacity in a particular availability zone, but it is locked to a specific size. **You can also buy a regional reservation. Now regional doesn't lock you to a particular availability zone, but it also doesn't reserve any capacity. Now if you use regional reservation, that also gives you some flexibility on size.** If you purchase a reservation for a T3.medium in a particular region, then you can also run two T3.smalls and have both benefit. **Regional reservations also benefit in part,** so if you buy a regional reservation for a T3.medium and then launch a larger T3 instance then you can get benefits on part of that instance. So a part reduction in hourly rate or a partial full reduction in the instance costs, but again that's only when you use regional reserved purchases. So when you buy these reservations, remember they're zonal or they're regional. **Zonal reserves capacity. Region gives you a lot more flexibility.**

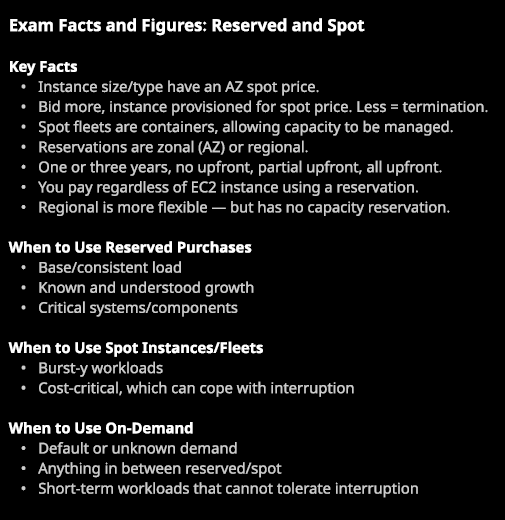
Now let's talk about capacity. **Capacity reservation is included when you do a zonal reservation, so when you pick a particular availability zone, you reserve capacity.** If you use regional, you don't and what does that mean? Well, imagine the AWS do have capacity constrains in a particular availability zone. The way that they resolve capacity constraints is they prioritize anyone who's reserved capacity. So anyone who's got zonal reserved instances have guaranteed access to that capacity. The second priority is on demand instances and the last priority, are spot instances. So if you've got any stringent requirements, where you need to make sure you absolutely can start up an instance in a particular availability zone. Then you need to make sure that you get that capacity.

Now for the Solutions Architect exam, you need to know when and where to use **reserved instances versus on demand versus spot.**

**Instances for a particular size and type in an AZ, they have a spot price. If you bid more than the spot price in a particular easy, then you get access to that instance for that spot price but if your maximum bid is ever exceeded by the spot price, you lose access to that instance. It gets terminated but you do get some warning.**

**Spot fleets are like containers. You can use different types and size together with on demand. They essentially allow you to manage capacity as one single container. Reservations are either zonal, so AZ scoped or they're regional. They're one or three years, and you either pay no upfront, part of it upfront, or all of it upfront. The key thing, though, is you pay for the EC2 instance regardless if you're using a reservation. So if you buy a reservation and you don't utilize it you waste that money. Now regional is more flexible, but it has no capacity reservations.** So keep that in mind, that is an important one for the exam. It does come up if you get any situations where you're running short on capacity, where you've got a requirement where you need to reserve that capacity then if you're using instance reservations it needs to be zonal reservations. So when to use reserved instances? Well, if you face exam questions, where you've got scenarios. We've got a base or a consistent load then use reserved instances. **If it's a known or understood growth, then you should use reserved instances. If it's critical systems or components where you need to absolutely ensure that they have the capacity inside an AZ and you want to achieve the cost benefits use reserved instances do not use reserved instances when it's bursty load.** When you don't understand that load, or where you could in theory, change the type and size of instances in the future or on a regular basis. You might be able to use a service that does support reserved instances, but also services might not.

So as I go through the course, I'll talk to you about which services do and don't support reserved instances. **When you'd use spot instances or spot fleets is when you've got bursty workloads. So if the workload is cost critical or they're massively variable then spot instances can fill that gap and do so cheaply, but only if you can cope with interruption.**



There are certain services and an example of this is **EMR or elastic map reduce** that I've not discussed with you yet where certain components of that can you spot instances and certain components you should not use spot instances for. So you need to have some knowledge is to go through the course of when it's appropriate to use spot instances for certain services but in general, **if it's bursty workloads, cost critical ,and you can cope with interruption or, for example, with web applications if the state is stored off the server, then yet you can use spot instances**.

Now when would you use on demand? Well, it's the default. You should start off thinking on demand and only moved to one of the other two billing models if you know that your usage fits**. If your demand is unknown or uncertain than use on demand and you can always look to move to spot or reserved later down the road as you get an overview of exactly what your workload is. On demand are great for anything between reserved and spot.**

**So if you've got a system that utilizes all of these three different billing models then on demand are great when your services can't tolerate interruptions and you just need to provision scaleable capacity**. **On demand are also good for short term work loads that cannot tolerate interruptions.** So when I'm creating instances for this course I'm not going to be utilizing spot instances because I don't want them to be terminated halfway through making a lesson. I'm generally using on demand to connect to, because I know that they're consistent and won't be interrupted throughout the course of a lesson. So if you've got any situations, any applications that absolutely can't tolerate interruption or they're short term you just need to spin up an instance or produce and work and then terminate it well then on demand is the best approach. Now you absolutely will get questions on this on the exam so spend the time thinking about any workloads within your current employer, previous employers, or other systems that you've got experience with and try to fit them into one of these three billing model categories. Try to work out for all the systems you could experience with exactly how you'd structure them from a cost billing model perspective and if you do that constantly as you go through the course.