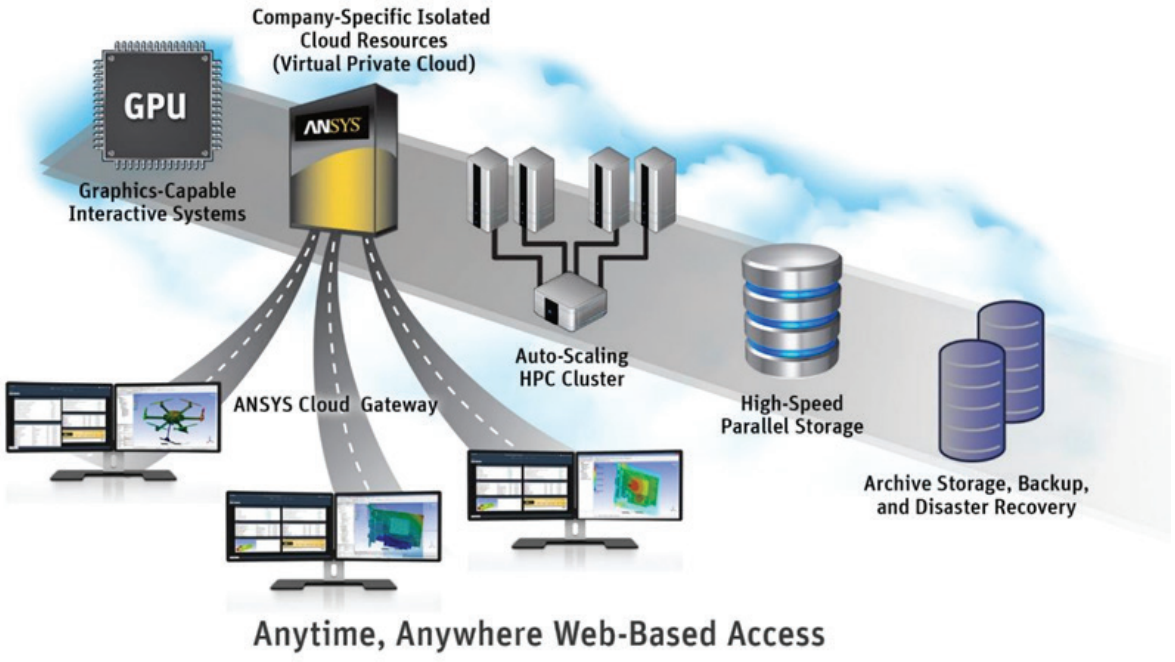
Application Brief

Cost Optimization for Cloud-Based Engineering Simulation Using ANSYS Enterprise Cloud

Most users of engineering simulation are constrained by computing resources to some degree. They are typically hungry for more resources to run ever more and larger simulations, and to solve their engineering problems faster to deliver better products to the market more quickly. Cloud computing infrastructure providers like Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform offer the promise of unprecedented scale, global availability and pricing models that allow companies to use the computing power they need, when they need it, and pay only for what they use. However, companies that are migrating to the cloud are often challenged to understand the complexities of cloud pricing models. Our goal with this paper is to provide a better understanding of cloud pricing models, and to offer some strategies for cost-optimization of the infrastructure portion of cloud-based engineering simulation performed within ANSYS Enterprise Cloud.

ANSYS Enterprise Cloud provides a turnkey environment for performing end-to-end engineering simulation running on AWS public cloud infrastructure. AWS Elastic Compute Cloud (EC2) provides the underlying infrastructure at a global scale, and ANSYS has made it easy to take advantage of the cloud as a platform that enables broader adoption of engineering simulation. By giving engineers access to limitless computing power from any location, ANSYS Enterprise Cloud can help drive product innovation and deliver significant business value.



Why should I consider the cloud as my simulation hardware platform?
The primary benefits of cloud infrastructure are:

- **Agility** — new computing infrastructure can be provisioned in minutes
- **Elasticity** — the amount of infrastructure in use can be varied in real-time in response to computational needs at the moment
- **Accessibility** — solutions can be deployed in data centers distributed globally, and can be accessed from anywhere using a variety of devices.

Companies in the early stages of cloud adoption often find themselves comparing the cost of these substantial benefits to the cost of updating or expanding on-premise resources. While this cost-centric viewpoint may be shortsighted (we feel that you would be better-served to focus on the value of overall simulation productivity), the business reality is that cost is a significant factor that must be considered.

How much will it cost?
When it comes to using the cloud, it's more challenging than you might think to predict, in advance, how much you will spend servicing your simulation computing needs. Since the cloud allows you to use the hardware you need, when you need it, and pay only for what you use, it should be clear that in order to answer the "how much will it cost?" question, you need to know what you plan to use.

As counterintuitive as this sounds, you should resist simplifying things in an attempt to compare "apples to apples." The temptation is to compare the cost of a server that's running full time on the cloud versus the cost of buying a similar server on-premise, but there are three problems with this comparison:

- By pricing a server running full time in the cloud, you've eliminated one of the primary benefits of the cloud — elasticity
- In considering the price of the on-premise server, you might neglect to consider the full price of the server. In addition to the purchase price, you need to consider staff for installation and provisioning, ongoing maintenance, power, cooling, and the cost of the data center space itself.
- It ignores the many intangible benefits of cloud outlined in the previous section.

How does cloud pricing work?
To further complicate matters, public cloud infrastructure providers like AWS, Microsoft Azure and Google Cloud Platform have several pricing models that offer flexibility and the opportunity for cost optimization to the customer. Our goal here is not to engage in an exhaustive discussion of the various pricing models offered, or to compare the pricing models among the various providers. Instead, here we hope to give you an understanding about what can be done to optimize your spending on hardware when it comes to performing engineering simulation on the cloud.

Since the solution we're focused on here is ANSYS Enterprise Cloud, which is currently only available on AWS, let's concentrate on the different pricing models available from AWS before considering how we might use them to optimize our overall cloud spend. AWS EC2 allows you to provision virtual machines (the term AWS uses for a virtual machine is an "instance") and bills you for those machines on an hourly basis. The hourly rate you pay for your instance depends on the type of instance used and the pricing model you select. AWS provides dozens of instance types, each optimized for a particular computing need. There are variations within each category, but instance pricing breaks down into three categories, as described below:

1. **On-demand instances.** This is the most flexible model where you provision the instance you need when you need it (on-demand) and pay a fixed hourly rate while you use it, with no long-term commitment. If you stop the instance, you no longer pay for it. This is the most expensive model on a per-hour basis. This pricing model is well-suited to use cases where infrastructure needs are highly variable.
2. **Reserved instances.** By committing to use a particular instance on a long-term basis you accomplish two things: you guarantee availability, and you get a lower per-hour price (the rate varies, but it's typically about a 40 percent savings on a per-hour basis versus on-demand). Different duration options are available, but the most commonly used requires an annual commitment. This pricing model is well-suited to use cases where instances need to be up and running constantly.
3. **Spot instances**. This is a very interesting model for HPC. Spot instances allow you to bid on spare EC2 computing capacity. This is a market-driven pricing model that varies depending on the amount of spare capacity in the cloud data center and the demand for that capacity. You establish the price you're willing to pay for a given instance (your bid price) and if the market price is below your bid price, you pay the market price. Of course, market-driven prices vary, but for the machines we use for HPC in ANSYS Enterprise Cloud, we've observed that the spot market price is typically about one quarter of the on-demand price, so clearly spot instances can result in significant cost savings. The catch is that while you're using your instance, if the market price goes up and exceeds your bid price, you lose your instance. There are strategies to manage such occurrences, but we will come back to them later.

Which pricing model should I use for ANSYS Enterprise Cloud?
Actually, all three pricing models can and should be used. ANSYS Enterprise Cloud is intended to be a persistent data center that's always available and ready to serve your simulation needs. As such, there are some instances which are running at all times — the so-called **fixed infrastructure**. Since this infrastructure needs to be running at all times, it makes sense to commit to using **reserved instances** to ensure availability and benefit from the lower hourly price.

For heavy use of simulation however, the fixed infrastructure tends to be a relatively smaller portion of your AWS bill. The total cost is dominated by the variable use of infrastructure that scales up and down on demand. There are two primary types of **variable infrastructure**:

1. **Interactive sessions**. These are instances used for interactive tasks like pre- and post-processing, where you work interactively with the software just like you would on your local workstation. Because these tend to be used only during a portion of your working hours, these are best served by **on-demand** pricing. Spot pricing would not work well here because it would be too disruptive if the instance were lost due to a spike in the spot price market while you were working interactively.
2. **HPC instances**. Most heavy-solve tasks are done by submitting the solution to execute on an HPC cluster in batch. On the cloud, our HPC cluster is auto-scaling: It starts the instances required for the solution when the job is submitted, and shuts them down once the job is finished. Obviously, this kind of task is well-suited to the cloud **on-demand** use and pricing model. The downside is that this can make cloud-based HPC jobs expensive when (carefully) compared with the price of executing the same job on on-premise infrastructure. With heavy use of HPC for production simulation workloads, the cost of running these HPC instances tends to be the dominant portion of your overall cloud hardware spend, since we're typically using the largest, fastest, compute-optimized instances that are available (remember, the "HP" in HPC stands for high-performance). This is where AWS **spot instances** can be applied effectively to lower your cloud spend, if used with some care. We'll cover some best practices on how to do that in the sections below.

So if spot pricing is low cost, why doesn't everyone use it?
AWS Spot instance pricing is probably used less than it should be because of the fear of spot termination. Even the term "spot termination" sounds dire. As mentioned previously, the advantage of spot instances is that they are typically one-quarter the cost on a per-hour basis when compared with on-demand instances. To be clear, you don't get inferior hardware when you use spot instances; you get the same machine, just at a lower price with the understanding that you're getting that price because you're using a machine that would otherwise be idle. The catch is that the spot price is a market-driven price that is determined by supply and demand for that unused cloud capacity. If the pool of idle instances shrinks, or if the demand for that pool increases, the market price could jump up above your bid price and you will lose your instance. This is referred to as "spot termination."

However, if you let the fear of spot termination deter you from using spot instances, you're missing an opportunity to lower the cost of your HPC by as much as a factor of four. There are two good reasons why you CAN effectively make use of spot pricing for HPC:

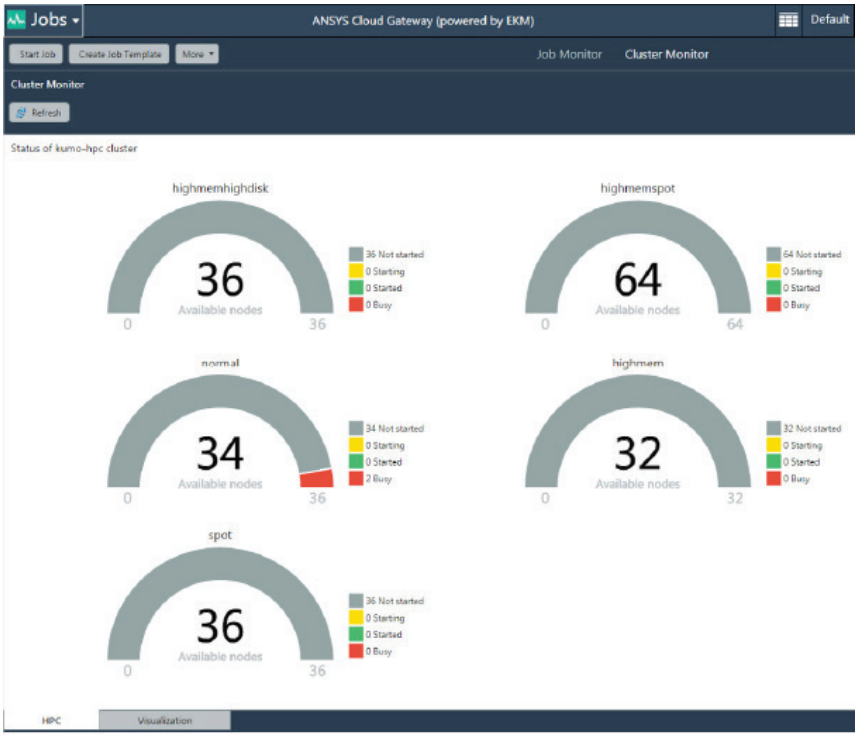
1. **The market is variable, but it's often quite stable.** The spot price is established by the supply of, and demand for, your desired instance in your desired **availability zone** (AZ) (the term AWS uses to describe a pool of compute resources in a particular region). In some cases, the spot price is quite stable. We have observed periods of weeks for some of our HPC instance types during which the spot price stayed well below the on-demand price. During periods like this, even very long-running jobs would run reliably, yielding valuable engineering results while using low-cost hardware.

2. **Spot termination does not have to cause loss of data or even significant loss of productivity.** A careful simulation engineer recognizes that a simulation job that is going to run for days or weeks should be set up to be recoverable in case of computer failure. The best practice in this case is to save intermediate results files so that the job can be recovered in the event of failure. When running using spot instances on the cloud, that same best practice is recommended. In ANSYS Enterprise Cloud, all files are written to work-in-progress storage that is based on Amazon's EBS (Elastic Block Store) service. This storage is not lost when a compute node goes down, so in the event of a failure (which could include spot termination) you can restart the job from the last-saved intermediate result, using on-demand instances, if necessary.

How do I use spot pricing for HPC with ANSYS Enterprise Cloud?
So far we have described the AWS instance pricing models and how to best use them. Now we will provide some specific hands-on instructions and recommendations for how to use spot pricing for HPC jobs in ANSYS Enterprise Cloud to best advantage.

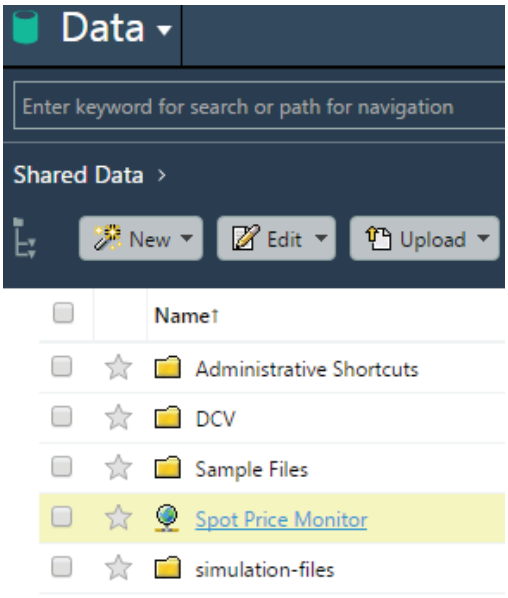
The first thing to note is that, when you run HPC jobs in ANSYS Enterprise Cloud, you explicitly choose whether to use on-demand instances or spot instances. You do this by choosing which queue you submit the job to.

If you check the HPC Cluster Monitor on the Jobs page of the ANSYS Cloud Gateway (an example is shown below), you should observe several queues. We use different queues for different solvers so that we can use the instance type best optimized for that solver in terms of core count, system memory and local storage. For a given solver, there will be two queues: one for on-demand instances and another for spot instances. For example, CFD jobs by default use a queue labeled "normal"; when jobs are submitted to this queue, they are run on high-performance compute-optimized instances at on-demand pricing. There is also a queue labeled "spot," which uses the same type of instances, but at spot pricing.



When running a batch job, or submitting a job to the HPC cluster from an interactive session, you'll need to explicitly choose one of these queues when you submit your job.

If the AWS spot market is variable, how will you know if you will actually GET the compute instance you request when you submit a job to a spot queue? The first factor is your bid price, the price you are willing to pay for that particular instance. If the current market price is lower than your bid price, you will get the compute instance when jobs are submitted and your job will run. Establishing the bid price is done by the administrator (we'll come back to bid price strategies shortly).

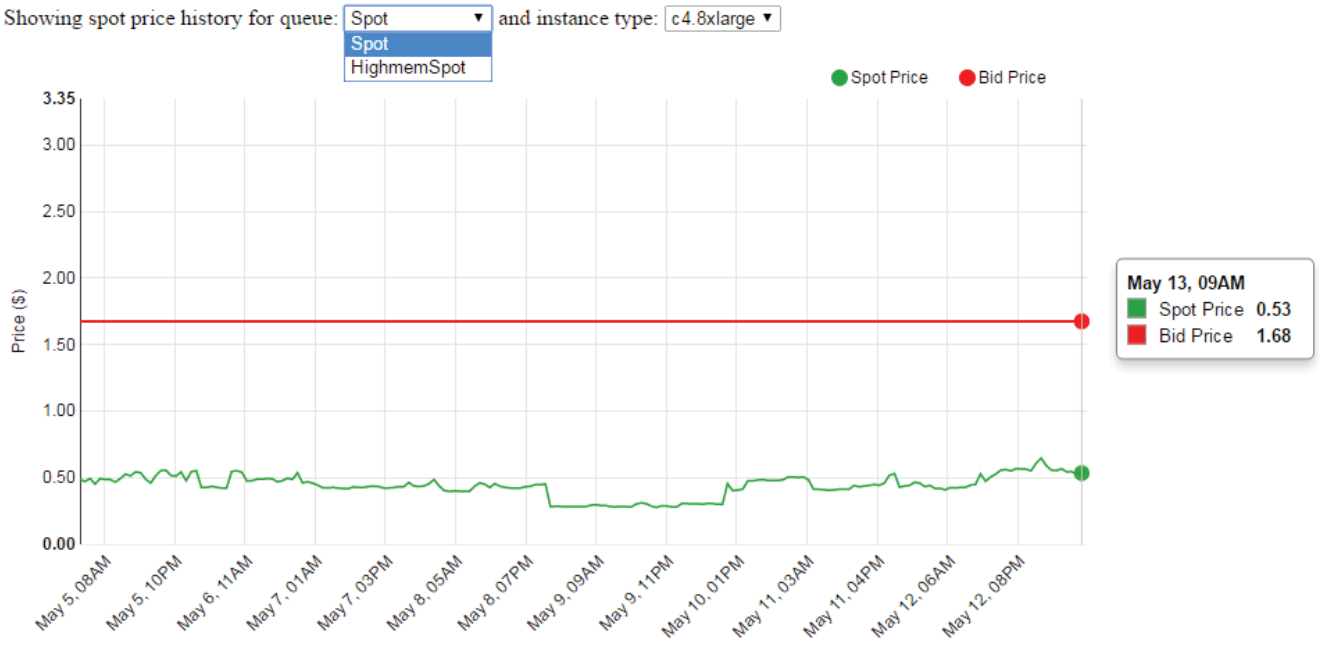


As a user who wants to run a job at lower spot market prices, it helps to know how the spot price market has been behaving recently for the instance type of interest. It is recommended that you check this BEFORE you submit your job. To do that, go to the root of the SharedData view on the Data page. There you will see a link called "Spot Price Monitor."

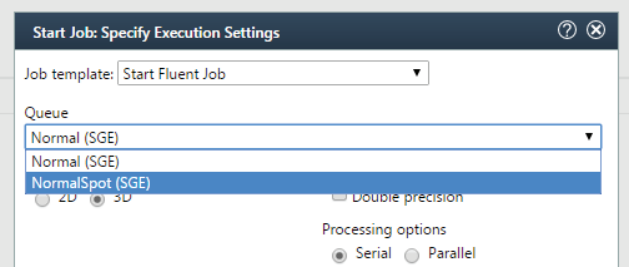
Click that link and after a short delay (during which the data is retrieved from AWS) you will see the recent price history, as shown in the example below:

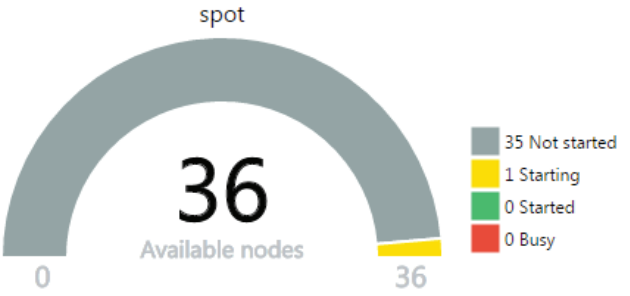
You can use the "queue" drop down menu to select the queue of interest. In the example above (which shows real data from May 2016), you can see that the spot price has been consistently below our bid price for several days (here we've chosen to set the bid price equal to the on-demand price, but this is configurable). As noted in the on-screen message, the risk of spot termination is therefore likely quite low. You could safely submit jobs to this queue without fear that your instance would be lost due to market volatility, and you'd be saving money on your job.

So with the knowledge that the sport market has been stable, we can proceed to submit our job to the HPC queue of interest. The queue selection is done at the time the HPC job is launched, either using a batch job template in the web UI (an example of the Fluent batch job template is shown below), or when submitting the solution to RSM from interactive sessions.



Risk of spot termination is low





Once the job is submitted, after about a minute an instance (or instances, if you're running a multinode parallel job) will show as "starting" in the HPC Cluster Monitor.

Spot instances take some extra time to provision, but otherwise they work the same as on-demand instances. If after 12 minutes or so your instance still shows as "starting" it may be that the spot market price has jumped above your bid price. If that happens, your job will remain queued until the market price drops below your bid price. You could go back to the Spot Price Monitor to check the market price again. If you can't afford to wait, you could cancel that job and resubmit to the normal (on-demand) queue.

What is a good spot price bidding strategy?

As mentioned earlier, there are two factors that influence whether or not you will get a spot instance when you submit a job to the spot queue: the market price and your bid price. If the market price is below your bid price, you will get your instance (you pay the market price, not your bid price), and you'll keep it for as long as the job runs and the market price remains below your bid price. The total cost that you pay for your job is effectively the hour-discretized integral of the market price versus time curve.

So this begs the question, "what is a good spot price bidding strategy?"

There is no single answer except to say that your bidding strategy should be based on your relative tolerance for two risks:

1. Set your bid price high and you run the risk of paying a lot for your job if the market price is high
2. Set your bid price low and you run the risk that you won't get your spot instance, or that instance might be terminated mid-run if the market price spikes up

A common practice is to set the spot bid price at or near the on-demand price for that instance. That way you have a reasonable probability of success, while not risking paying the on-demand price. Some customers big higher than the on-demand price knowing that, since they pay the market price, the price they pay will typically average out to be less than the on-demand price since market price spikes tend to be short-lived, and bidding a higher than on-demand price increases the probability of success with jobs submitted to spot instances.

Regardless of the bidding strategy you use, it makes sense for administrators to keep an eye on spending patterns to be sure that the practices they are using are working well and yielding overall cost savings.

What about software licensing?
This document focuses primarily on the cloud-based hardware for your simulation tasks and how you can optimize your AWS spend. Of course, hardware is not the only cost of performing simulation: there is the cost of software licensing to consider as well. Traditional software lease and paid-up licenses were really designed to serve the needs of fixed infrastructure — a set of workstations or an HPC cluster of fixed size. To address this, ANSYS has recently announced a new pay-per-use licensing model called ANSYS Elastic Licensing. Like the ANSYS Enterprise Cloud solution, Elastic Licensing provides access to the full ANSYS software portfolio, and allows you to use what you need when you need it, and pay only for what you use. Like the cloud hardware model, this introduces new possibilities for how you might optimize your overall spend on simulation to yield maximum engineering benefit. More details on Elastic Licensing can be found in this blog.

Some Closing Thoughts
This document has provided an overview of cloud pricing models and a high-level description of how they could be applied to performing engineering simulation using ANSYS Enterprise Cloud, while optimizing your hardware spend on AWS. It is our hope that, used efficiently, the cloud will provide a platform that, by virtue of its inherent accessibility and scalability, helps drive engineering simulation to yield better products that are brought to market more quickly.

ANSYS, Inc.
Southpointe
2600 ANSYS Drive
Canonsburg, PA 15317
U.S.A.
724.746.3304
ansysinfo@ansys.com

If you've ever seen a rocket launch, flown on an airplane, driven a car, used a computer, touched a mobile device, crossed a bridge or put on wearable technology, chances are you've used a product where ANSYS software played a critical role in its creation. ANSYS is the global leader in engineering simulation. We help the world's most innovative companies deliver radically better products to their customers. By offering the best and broadest portfolio of engineering simulation software, we help them solve the most complex design challenges and engineer products limited only by imagination.
Visit www.ansys.com for more information.