Real experts.
Real data.
Real savings.

WHITE PAPER

The Definitive IT Sourcing Guide to Cloud Vendor Competitiveness, Pricing and Support

PART II: COMPARISON OF CLOUD VENDOR PRICING

Featuring comparisons among Amazon Web Services, Microsoft, Google, IBM and Oracle

Table of Contents

About This Series

The cloud may be pervasive, but doing business with cloud vendors (particularly IaaS/PaaS) is still immature in many ways. What are each vendor's strengths and weaknesses? What pricing nuances should be considered? And how does support compare?

From an IT Sourcing perspective, this three-part white paper series explores what really sets five IaaS/PaaS vendors apart in the following three areas:

Part I: Competitiveness

Part II: Pricing

Part III: Support

Pricing for enterprise IT offerings is complicated, opaque and inconsistent – and the IaaS and PaaS categories are no exception. While we gave a nod to certain provider pricing pros and cons in the first part of this white paper series (***Part I – A Pragmatic Review of Cloud Vendor Strengths and Weaknesses***), it's impossible to designate any provider as "cheapest" or "most expensive." Cost comparison of a cloud service is far too complex and far too dependent on a host of variables and circumstances. The cheapest solution will not always be from one "low cost" provider and will depend heavily on the characteristics of your specific workload. To illustrate, let's look at some examples.

Configuration: Small Windows Config with Concentrated Workload

Figure 1 charts pricing for the major cloud service providers against hourly usage for a relatively small configuration running Windows (4 VMs with 2 cores and 4 GB RAM, 2 VMs with 4 cores and 8 GB RAM and 1 TB of block storage). Looking at usage is mandatory any time cloud pricing is compared, since one of the key promises of cloud is that you save money by paying only for what you use. This is an accurate claim… sometimes. It's impossible to confirm until you actually model costs at certain usage levels.

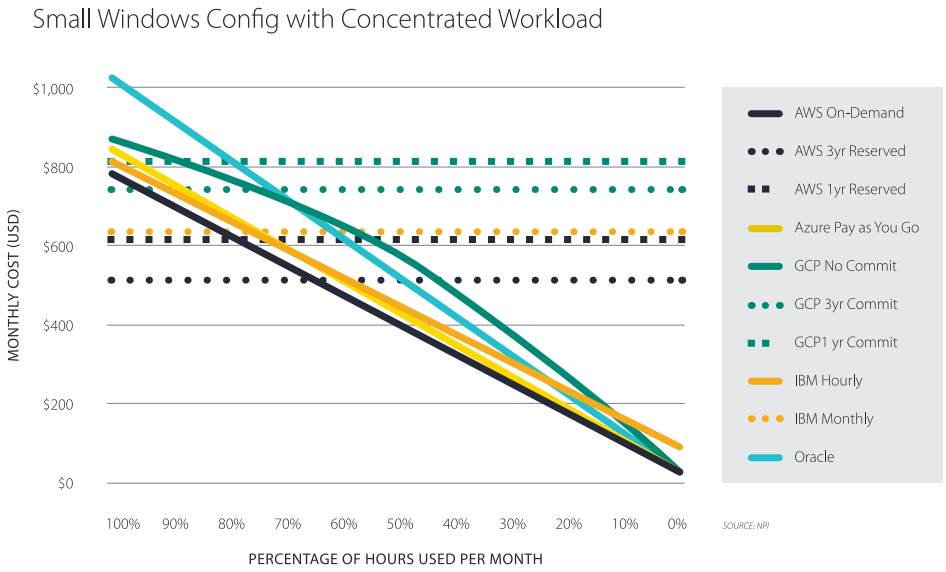
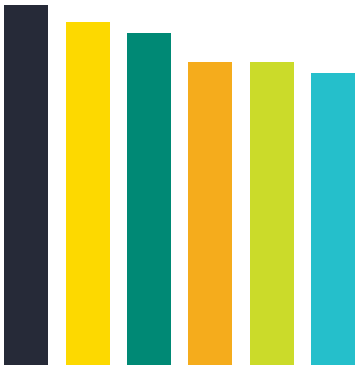


Figure 1

In Figure 1, the percentage of hours used appears on the horizontal axis, with full utilization on the left and tapering off to zero hours used on the far right. For now, we're assuming a "concentrated" workload, so 50% of hours used means 12 hours per day and not half an hour during each hour of the day. It's important to note this has a significant impact, and we'll discuss that in detail in later examples.

Dotted lines indicate purchasing options that require commitments to continue using the infrastructure for a period of time. In these cases, the customer gives up the pay-for-use pricing model in exchange for a lower, flat price. At the time this analysis was performed for this particular configuration, flat pricing is only offered by AWS, Google and IBM. Microsoft doesn't yet offer reserved instances for their A-series configurations, and though Oracle has announced flat monthly pricing, it isn't available yet. If you want to pay flat rates for Oracle cloud today, you'll probably have to negotiate the pricing directly.

According to a survey of more than 624 IT and Sourcing professionals, respondents reported IaaS/PaaS usage by vendor ranked by most to least annual spend as:



SOURCE: NPI CLIENT SURVEY, APRIL 2018

- Amazon Web Services
- Microsoft
- Google
- IBM
- Other
- Oracle

OBSERVATIONS

- One of the first things you notice when looking at Figure 1 is that GCP's (Google) on-demand pricing is represented by a gently curving line while all the others are straight. That's because Google builds something called a "sustained use" discount into its GCP pricing that increases as your usage increases. Originally this was probably designed to be an improvement on AWS reserved instances so customers wouldn't have to change instance types or payment terms just because a workload increased, thus simplifying the financial management of the solution. Apparently, that wasn't a good enough plan, either in terms of what customers demanded or profitability, because Google has since added 1-year and 3-year commitment offerings of its own, but the sustained use discount mechanism still remains. It also means that for on-demand, no-commit GCP instances, pricing tends to be best at high usage levels. Indeed, the already-discounted price is the one quoted in its pricing calculator, thus putting GCP's best foot forward when you're ordering the service.
- AWS is the cheapest option here if you're using On-Demand, but Azure and IBM SoftLayer (IBM Cloud) are very close. AWS has a clear edge over GCP for reserved instance pricing on this configuration.
- Something very useful that you can do with a chart like this is to identify the break-even point for pay-as-you-go (or on-demand) instances versus "reserved" style instances that have 1- or 3-year commitments. Reserved instances are always cheaper if you're fully utilizing them, and on-demand instances are cheaper for low usage levels. Knowing where the break-even point is will allow you to optimize your savings, assuming of course that you know what your workload will be or can manage your instance types dynamically after the application is up and running. If you're an AWS customer, we can see from the chart that 3-year reserved instances stop being cheaper at around 65% of hours used. Your application will need to release the instances for about 8.4 hours per day to reach that point. If it can release them even more than that, you'll save more and more money by using on-demand instances. For GCP, that happens at around 70% of hours used, and for IBM, it's at 73%.

Price Ranges: Small Windows Configuration with Concentrated Workload

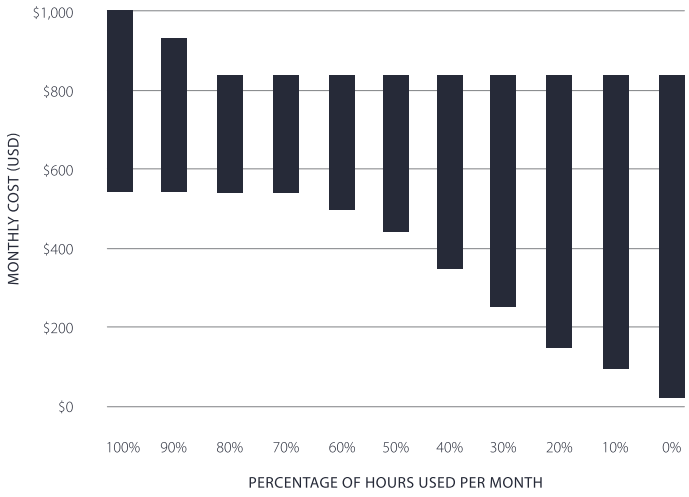
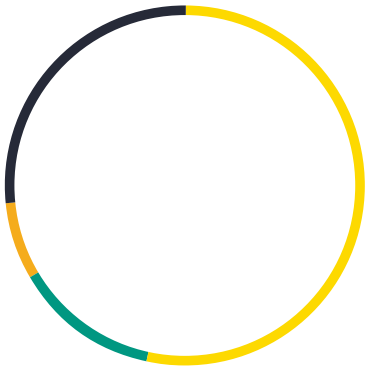


Figure 2

How would you describe doing business with IaaS/PaaS providers?



SOURCE: NPI CLIENT SURVEY, APRIL 2018

- 26.7% Good and getting better
- 53.3% Okay, but improving
- 13.3% Okay, but not as easy as it used to be
- 6.7 % Tough and getting worse

- Oracle's pricing starts out the highest of the group at 100% of hours used, where it's about double the price of an AWS reserved instance. The reason for this is clear: Oracle doesn't have configurations small enough for an apples-to-apples comparison here. In order to get the number of processors we needed on the VMs, we had to significantly over-provision the memory on the Oracle instances, so we're paying for that unused capacity.
- Given the analysis above, the cost for this configuration varies widely. Depending on your usage levels, you could spend anywhere between 1.5 and 20 times as much as you need to, based on the ratio between the highest and the lowest price. That's the top of the bar in Figure 2 versus the bottom. The length of each bar is how much money you could leave on the table if your cloud services aren't carefully selected, configured and managed.

> The truth is that we're still paying for access to capacity – not use of it. We're just paying in smaller chunks.

Configuration: Large Linux Config with Concentrated Workload

For this next configuration, let's replace the Windows requirement with basic Linux, which is a no-charge option for all of these services. We'll also go with bigger, more memory-intensive configurations that fit Oracle's standard options better, and ten times the amount of disk storage. That requires us to move to "memory optimized" instances, such as AWS's "r4", GCP's "highmem" and Microsoft's "A2m" types (or D and E types for Reserved instances, since the A series isn't available as reserved). Notice how different Figure 3 is from Figure 1:

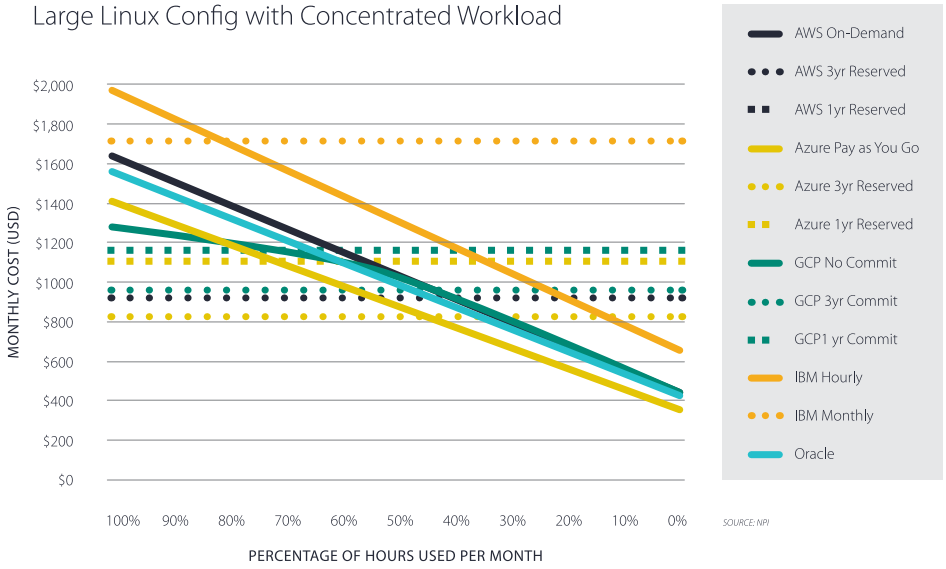


Figure 3. Note: GCP and AWS 1-year options are so close that Amazon's 1-year line in the chart is invisible beneath GCP's.

OBSERVATIONS

- IBM is significantly higher than the others, due to a higher price for storage and for more memory-intensive VMs.
- If you want to pay only for what you use, GCP starts out the cheapest but is overtaken by Azure at about 77% of hours used. That's important, because reserved instances are even cheaper than GCP's on-demand price above that usage level. For all usage levels below that, Azure (3-year Reserved) is the cheapest on-demand service.

- Reserved instance pricing is much closer between AWS and GCP for this configuration, and Azure comes in a bit lower than both. GCP and AWS 1-year options are so close that Amazon's dotted line in the chart is invisible beneath GCP's 1-year dotted line.
- The break-even point for reserved instances has also shifted downward. For AWS, you should now switch from 3-year reserved instances to on-demand at about 40% of hours used per month. GCP customers should switch away from 3-year commitments at around 46%. For Azure it's 45%. IBM's break point, on the other hand, shifted upward to about 80% of hours used.

This is a good time to point out the realities of usage-based pricing – including per-minute pricing options. The first thing to understand is that compute instance pricing from the leading IaaS providers is not really usage-based. If you spin up an on-demand compute instance and don't release it for an hour, but your workload only uses 5% of the CPU's horsepower for that hour, you will *still* pay the same price as if it used 100%. Disappointing, right?

The marketing hype in the early days of cloud was so good that this fact still comes as a shock to a lot of people. Yes, we're getting *closer* to true usage-based pricing than when we were paying by the month on a standard 5-year outsourcing contract. But the truth is that we're still paying for access to capacity – not use of it. We're just paying in smaller chunks.

(NOTE: AWS introduced per-second billing for some compute instances in October 2017. While this is a step in the right direction, there are numerous limitations and it's not available for all AWS services.)

While this only matters for applications with certain types of workloads, it's still a serious cost consideration. When it does matter, it can matter *a lot*. Figures 1 and 3 assumed that workloads are as concentrated into the smallest number of hours possible, so that 50% usage means 12 hours out of every 24. Even with hourly pricing that's fine, because you won't be charged for the 12 hours each day that you aren't using.

But what if your workload is spread out? What if 50% means 30 minutes of usage each hour? With hourly pricing you pay for a full hour when you use any fraction of it, so you're going to get charged for the *full 24 hours*. With per-minute pricing you'll only be charged for 12. You could be paying *twice* with hourly what you'd pay with a per-minute pricing scheme.

What if your usage is 33%, and the usage is dispersed widely so that it translates to 20 minutes each hour of actual use? Now you'll pay *3 times more* with hourly pricing than with per-minute. That's exactly what will happen if you're using AWS On-Demand with Windows, IBM Cloud hourly pricing or Oracle. The more frequently your application releases its instances, the bigger an impact this has. This is a huge part of the reason that providers are making so much money on cloud. Every minute you aren't using an instance is a minute they can rent it to another customer, and you could still be paying for it until the hour is up, so they get to charge two customers for the same resources for that period of time.

The "realities" of cloud pricing illustrate a critical point in cloud cost management. As much as cloud providers try to offer the best mix of provisioning, it's not up to them to save the customer money. Cost savings are wholly the customer's responsibility.

If customers want to optimize cloud costs, they need to right-size instances, auto-scale to meet spikes in demand, scale back during non-peak hours and auto-park instances and/or environments when not in use. Alternatives are to use platform services that charge based on the number of transactions (serverless) or make better use of SPOT market instances.

Configuration: Small Windows Config with Dispersed Workload

Figure 4 takes the same configuration used in Figure 1 and changes the basic assumption to a workload that is dispersed over all hours of every day. With that assumption, all of the cost benefits of pay-per-use plans are completely erased if pricing is per-hour. Providers with per-minute pricing, however, can still deliver the same benefit of prices that decline as usage declines, since releasing an instance after only 1 minute of usage suspends charges.

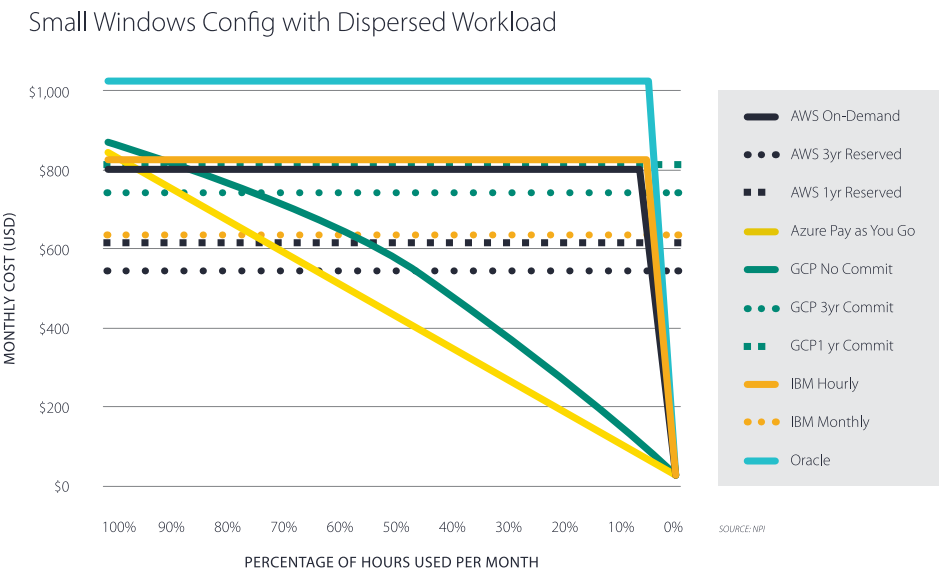


Figure 4

OBSERVATIONS

- AWS On-Demand, IBM Hourly and Oracle are no longer competitive at all because they don't have per-minute pricing (remember this is a Windows configuration, and AWS only offers per-minute pricing for Linux).
- Azure has the best pricing of any option below about 60% of minutes used (about 36 minutes per hour.
- Above that usage level, you'd be best off with an AWS 3-year reserved instance.
- If you're an AWS customer with Windows applications or an IBM customer, you should always use reserved instances or IBM's monthly pricing when workloads are highly dispersed over time.
- If you're a GCP customer with Windows applications, to save the most money you should still switch from 3-year commitments to no commitment at about 70% of minutes used for highly dispersed workloads, the same break-point as we saw for concentrated workloads.

Note that we are *not* assuming that the workload is dispersed across every available *minute*, as that would erase some of the benefits of per-minute pricing as well. That would be a relatively unusual situation, and it's fairly safe to assume that if the workload is idle one second and active the next that you should probably be looking at Function-as-a-Service (FaaS, i.e., "serverless") or possibly a reserved instance, depending on the volume of transactions.

Configuration: Large Linux Config with Dispersed Workload

For the large Linux configuration, changing the assumption to a dispersed workload across every available hour has a similar effect as it did with the small Windows configuration. The exception is AWS, since it *does* offer per-minute pricing for Linux.

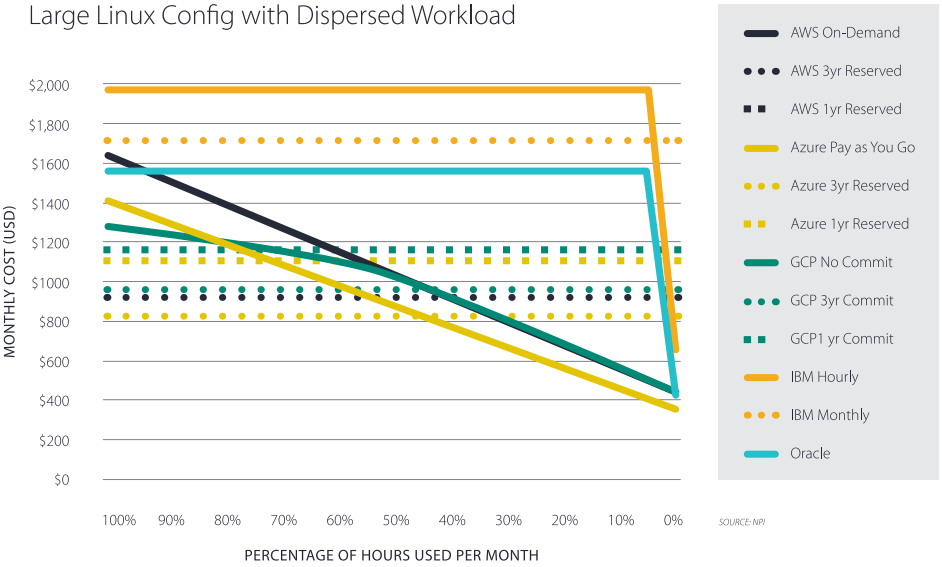


Figure 5

OBSERVATIONS

- IBM and Oracle remain uncompetitive due to their lack of per-minute pricing. Since this is a larger configuration, chosen to be in line with Oracle's configuration offerings, they now have a significantly lower price than IBM does.
- Usage levels make a big difference regarding which service is the most economical. For on-demand, pay-per-use services, GCP is cheapest above about 75% usage (though you'd be better off with a reserved instance at that level), but below that Azure has the pricing to beat, with GCP rapidly rising to reach parity with AWS as usage levels go down. At 100% usage, AWS On-Demand is $341 more expensive per month than No-Commitment GCP, which is a 27% premium over the GCP price. Based on this initial research and configuration comparison, Azure *appears* to be the safest choice for saving money here. However, it's worth noting that NPI's in-the-trenches experience with Azure deal negotiations indicates that Azure pricing is still cryptic based on the customer's unique requirements and the thousands of SKUs in play.
- 3-year reserved instances from Microsoft, Amazon or GCP are more economical than all of the On-Demand pricing models between 45% and 100% usage. So, if your needs are long-term, that's the easiest choice to make by far.

Additional Considerations

As mentioned earlier, comparing cloud pricing is a difficult task. Even in our attempts to normalize and create an apples-to-apples comparison, there are numerous variables that change customer to customer. Here are few considerations to keep in mind:

- Your mileage will vary. These are very simple configurations, and yours will likely be different. Price them out. Each application you put into the cloud will have a different cost profile. It's *understanding why* these costs are what they are that will save you the most money. You may even find that not all services are even *capable* of running your configurations, as not all versions of operating systems, databases and middleware are supported by every provider.
- Windows and Red Hat cost more than basic Linux builds like CentOS, Ubuntu and Debian. Windows can be a lot more, so make sure you're getting good benefits from your choice of OS.
- It's more than servers and storage, even though that's all we've modeled here. Don't forget about support, which each provider prices differently, though AWS, Azure and GCP do have similar tiered pricing structures. The significance of that component will depend on your monthly spend with the provider. You'll also need network bandwidth, and possibly database services, load balancing, VPN and more. Each of those has a separate cost.
- Performance matters. Not all providers will give you the same response times, and it can vary from one data center to the next. If you need to throw more hardware at your deployment to get the performance you need, that's a very real cost.
- Know what your workload will look like before you put it in the cloud. If you haven't modeled the configuration and the workload, you have no idea what your cloud service will cost, and that means you don't yet have a business case.
- AWS has the most options as far as how to pay for a reserved instance. You can pay it all up front, some up front or nothing up front. The middle option is used for this comparison, so your price could be a little higher or lower depending on how you pay.
- Microsoft has an additional charge of $0.0005 per 100,000 transactions for its standard magnetic storage, so actual prices will be a bit higher than what's shown.
- We've only looked at the most concentrated and most dispersed theoretical workloads here. Yours could easily be something in between those two scenarios.
- Look at the discounts available to you, especially if you're a Microsoft or Oracle customer. Below are published examples – but keep in mind that it's not comprehensive and providers may exhibit flexibility based on your unique computing needs and negotiation prowess.

Beware of these pricing tactices

Oracle: The cloud has triggered interesting behavior from Oracle. One example is allowing customers to redirect their on-premise maintenance spend to cloud spend. Less enticing, however, is the issue of re-pricing. In the past, Oracle re-priced certain orders when a customer tried to drop one line item (e.g. an underutilized database license) as a way to prevent revenue loss. Now Oracle sometimes allows customers to drop line items without re-pricing – but only if they replace that order component with a cloud service equivalent.

Microsoft: Nearly every Azure deal includes a Monetary Commit SKU. This is, in essence, a cloud spend commitment rather than a solution-driven SKU that's attached to specific customer requirements. This is a sure way to overspend given the fact that both Microsoft and many of its customers still struggle to spec their cloud requirements.

AWS: It's true that AWS is mostly transparent about its pricing and discounts. However, AWS is beginning to show some flexibility in discounts as well as the willingness to offer these discounts without a prepaid commitment.

PROVIDER	DISCOUNT
AWS	• Volume discounts for reserved instances in certain tiers: 5%, 10%, negotiated. • "Scheduled Reserved Instance Pricing" is actually a short-term reservation of an On-Demand instance, for certain regions and instance types, giving you a 5-10% reduction in the On-Demand price in return for scheduling them ahead of time on an hourly, daily or monthly basis. For example, if you know you'll need one day of processing on the first day of each month, this is for you.
Microsoft	• 5% prepayment discount on $6000 spend or more. • EA customers can use the "pre-purchase plan" to pay up-front with true-ups for growth at year-end. Negotiated discounts guarantee pricing "will not be beaten by AWS." Amount depends on commitment to minimum spend. The risk is in overcommitting to get a bigger discount and paying up front for Azure capacity that doesn't get used. • Temporary discounts for buying Azure as an "add-on" to a Windows Server CIS (annuity) license ("Compute Option)." • Software Assurance customers can use "license mobility" to essentially use Windows Server licenses on Azure without paying for Windows again. These licenses can still be used locally. Also referred to as the "hybrid benefit," and can be up to 40% savings on published Azure pricing, according to Microsoft.
Oracle	• BYOL (license). Bring your own DBMS, Analytics or Middleware license and Oracle won't charge you again in the cloud. • Universal Credits. This is a discount for up-front payments, similar to Microsoft's EA discounts for Azure. • Negotiated. That includes opportunities to save on your non-cloud spend in return for using cloud, e.g., Oracle software licenses. Savings can be substantial.
IBM	• IBM is restructuring account types into "Pay as You Go" and "Subscription," and there are discounts for committing to monthly minimums with Subscription accounts. There are also "Dedicated" accounts with minimum 1-year terms that carry discounts. • Negotiated. That includes opportunities to save on your non-cloud spend in return for using cloud, e.g., IBM data center contracts. Savings can be substantial.

Figure 6

The Four Pillars of Cloud Savings

As with most subscription services, there are four areas where savings can be derived:

1 Contract Optimization – This is optimization of pricing/rates, discounts and business terms. It's imperative that enterprises conduct price benchmark analysis to determine fair market value targets that they can then negotiate towards. Remember, cloud providers are beginning to show more negotiation flexibility. If you're not optimizing your contract (and pricing), you're leaving money on the table.

2 Service Optimization – Service optimization is where the most savings can occur. This is the managing of service utilization to exactly match your requirements. Examples include account consolidation (making sure you have all accounts under one bill payer so you can receive the appropriate discounts) as well as selecting the right services (e.g. reserved instances vs. on-demand pricing).

3 Audit for Contract and Pricing Compliance – Billing errors aren't as frequent with cloud providers as they are with other comparable types of usage-based vendors (for example, wireless carriers). However, industry best practice is to implement a process for reviewing and disputing, where appropriate, billing errors.

4 Demand Management – Industry best practice for managing demand is through establishing and enforcing tagging standards and baseline configurations as well as standardizing on a preferred instance type. This will help prevent out-of-control usage and cost spikes.

When asked if they could change anything about their experience with IaaS/PaaS providers, respondents expressed concern around pricing and support transparency as well as providers' ability to be a strategic partner in solving critical business problems. Sample responses include:

"Pricing, add-on costs, technical support, and better alignment to organization's needs instead of the vendor's."

"A more strategic approach to helping us solve business problems."

"Cost is still more expensive than on-premise."

"Ability to understand the full cost picture, including the IaaS/PaaS services and wrap-around services."

Source: NPI Client Survey, April 2018