AZURE COST MANAGEMENT: BACKGROUND, TOOLS, AND APPROACHES

As enterprises are transitioning to the cloud, it is becoming evident that cost management and computing are inexorably tied. This has always been true but has never been more evident than it has as the computing world embraces a “Pay-for-play” approach to computing, where there is a direct relationship between computing power and cost. The job of the CIO and CTO has changed from estimating annual budgets to monitoring spending and taking proactive steps to make good cost decisions.

This whitepaper explores an approach to cost management that spans from an overview of subscription choices down to how to approach coding in a cost-conscious manner and what tools to use. While it only touches on each subject area, it exposes the reader to concepts and practices that may not have been considered in the past and can be a starting point to launch into a deeper examination of each subject.

Much of the knowledge and concepts in this whitepaper come from AIS’s “in-the-field” work optimizing cloud usage as part of other engagements. Cost management is seldom the thrust of an engagement but is constantly considered in any cloud endeavor.

MOTIVATIONS FOR MOVING TO THE CLOUD

In any cloud cost management discussion, it is important to examine an organization's motivations for moving to the cloud. Cost reduction is not always the reason, and in fact may not be the best reason to move to the cloud. If you are only moving to save costs, it may well end up costing you more. Rather, it is about improving your time-to-value. Implementing new functionality, optimization of existing products, improving testing and quality and doing real load testing all decrease time-to-value. In today's rapidly changing technology landscape, it is crucial to an enterprise that Information Technology keeps pace with customer or citizen demand and expectations.

The value of cloud computing lies in the time-to-value computation. While immediate capital costs may not be reduced, a corporation will be able to keep pace with the market and realize a greater profit when they can respond to market pressures quickly. Similarly, government agencies will be ready to better serve citizens by presenting information in a familiar, modern way.

Finally, "**Cost Maturity**" will come with cloud maturity. As an organization moves away from compute and storage to things like containers and serverless computing, the enterprise will be able to make better and more informed decisions that over time will allow them to reduce costs over that of on-premises servers. Similar to the term "Cloud Maturity," we use the term "Cost Maturity" to describe this journey.

Part of this maturity is realizing how differently cloud computing affects costs and how to have those conversations with business users. This is a large and complex subject that is covered in excellent detail by Joe Weinman in his "10 Laws of Cloudonomics," which has recently been expanded upon and published in his book "Cloudonomics."

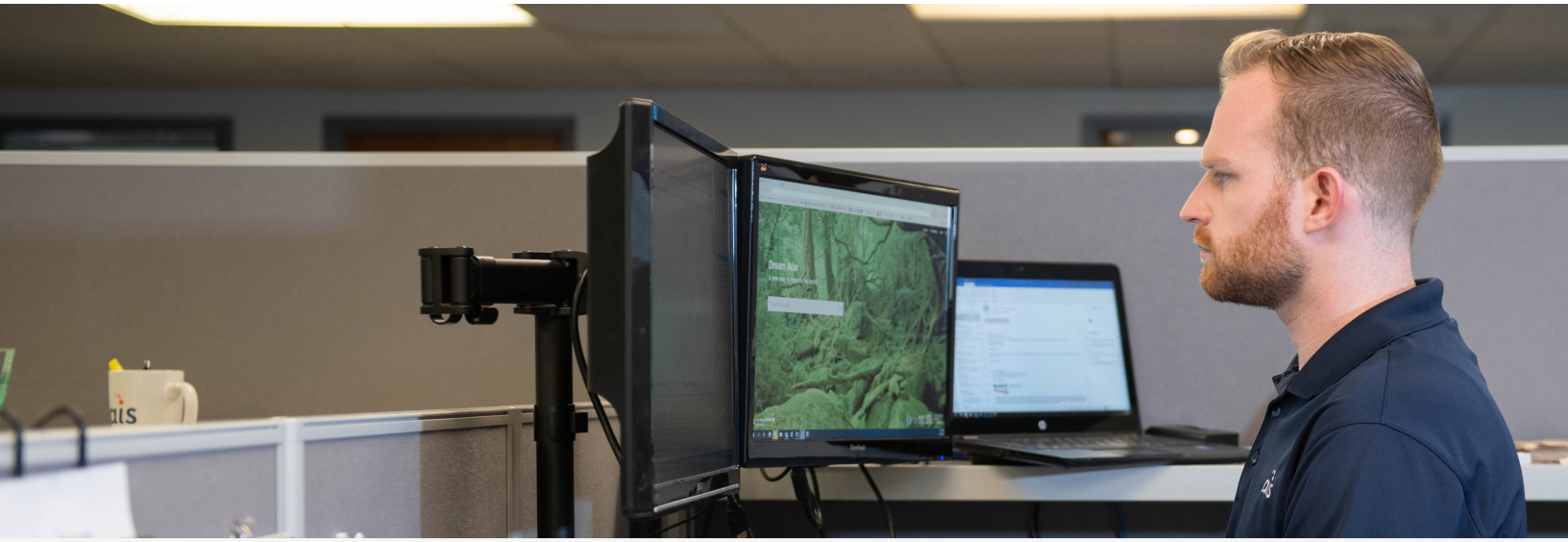
"REAL" ON-PREMISES COSTS

While the greater value of cloud is in agility and maturity rather than direct cost savings, often cost analysis will be flawed and disproportionately favor on-premises data centers. This is because cost analyses are often not detailed enough.

Does your cost analysis include rack cost, administration cost, salaries, depreciation, backup storage, air conditioning and power, rack spaces, and more? An enterprise wishing to perform an accurate cost analysis will need to determine all the calculations necessary, not just the cost of the servers and depreciation. AWS Total Cost of Ownership (TCO) Calculator is a great tool for such an analysis.

AIS recently performed this cost analysis for a large company that used several data centers. One sample calculation is for the cost per terabyte of information. This calculation included data center cost per square foot divided by the rack size, divided by servers/racks and virtual servers per server. Next, we added in similarly calculated administrator salaries, tools for troubleshooting and estimated outage escalations based on historical data, not to mention depreciation and refurbishing costs, which were missing from the initial cost estimate.

Once these "real" on-premises costs were calculated, we arrived at an "apples to apples," comparison. As a result, the flawed initial estimate — that moving to Azure would balloon costs — became a nearly equal comparison. With this, the customer was more comfortable paying for the advantages of the cloud. Furthermore, armed with this information, the customer was able to answer the question "how much is the agility, dependability, and time-to-value worth to us?"



Cost Optimization Types

Now that you're ready to move to the cloud, how can you improve your cost maturity? There are 3 types of cost optimization:

1. **Resource-based**
2. **Usage-based**
3. **Pricing-based**

1. RESOURCE-BASED OPTIMIZATION

In many scenarios, enterprises will separate cost discussion from a technical discussion to the extent that the cost and technical people never speak. It's the role of the cloud broker, cloud evangelist, CTO, or whoever controls access to cloud resources to have a conversation with the development team around "what is the right SKU?". Developers don't always have the tools or training to right-size an instance for an application. Out of caution, they often select one that is larger and more powerful than necessary.

Microservices-based architectures that afford higher density deployment can lead to the better utilization of resources. For example, Kubernetes scheduler uses a bin-packing algorithm to optimally place pods on cluster nodes. This can result in significant savings compared to VMs where the utilization tends to be low.

Planning along with developers or maintenance engineers will help to create a dialog that will save the enterprise money while maintaining the appropriate amount of reliability for the applications. This is one important steps towards cloud cost maturity.

2. USAGE-BASED OPTIMIZATION

Auto-scaling is a tool that can help with right-sizing but is not a "silver bullet." Auto-scaling is only as effective as the metrics used to perform the scaling. In AIS's experience, the best indicators for auto-scaling come from metrics reported by the application. Metrics such as queue length, page response time, etc., can be the best guides but must be discovered with the developers or maintainers of the system hosted. Again, this is a good indicator of an enterprise's cloud economic maturity.

One of the largest "leaks" of compute time is idle resources such as idle dev and test instances that are unused but maintained and available for beak-fix or feature additions. Typically when these systems are required, it's under urgent circumstances and off-hours. Hence, the systems remain idle given the time needed to spin up instances is a roadblock to the agility required to respond to a situation quickly.

A mature organization needs a plan for these instances and must have the correct dialog with developers to determine an approach to quickly re-provision dev and test servers. This should be a part of the DevOps pipeline and can be a useful application of Infrastructure as Code (IaC) within that pipeline.

3. PRICING-BASED OPTIMIZATION

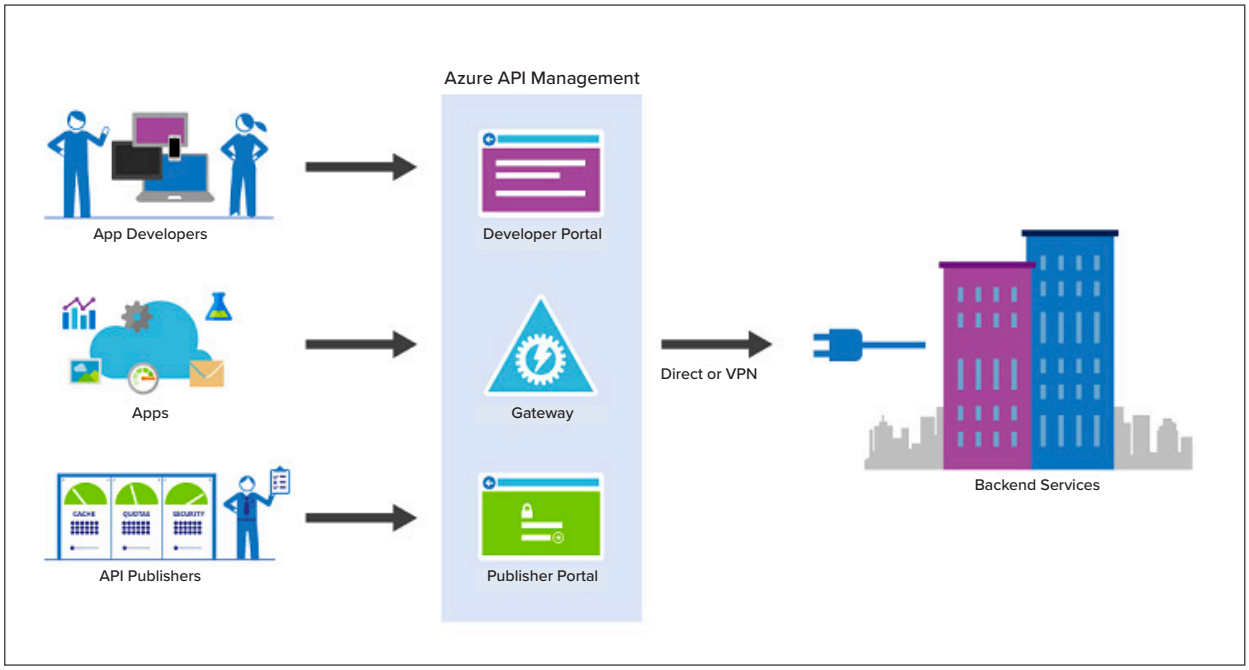
Reserved Instances are an excellent (and in Azure, simple) choice for workloads that the enterprise knows will always be running and must always be available, such as the corporate website or the main corporate application. Organizations that identify these applications and uses early will realize significant cost reduction by reserving instances for them. Reserved Instances make it easy to calculate the annual cost for computing and allow an organization to budget accordingly. Please refer to our blog on reserved instances for more information.

Tools and Governance

Shared Services are great for cost savings, but how do you push those costs down to the business units that use the shared service? Again, this is part of the cloud maturity, but Azure API can be used to create tools to tag and track use of share services in order to perform a chargeback as part of the normal billing process.

The idled instances mentioned in the cost optimization section above can be addressed in a variety of ways. The first issue is identifying those idle instances. An often-used rule of thumb for determining if an instance is “idled” is to check if CPU, Disk IO, and Network IO are less than 2% over eight hours (2% is to allow for usual health checks and other idle system activity). This is a point where governance overlaps with both operations and cost management. As part of the governance plan, an organization must proactively audit the provisioned resources to identify instances that are not compliant with established policies.

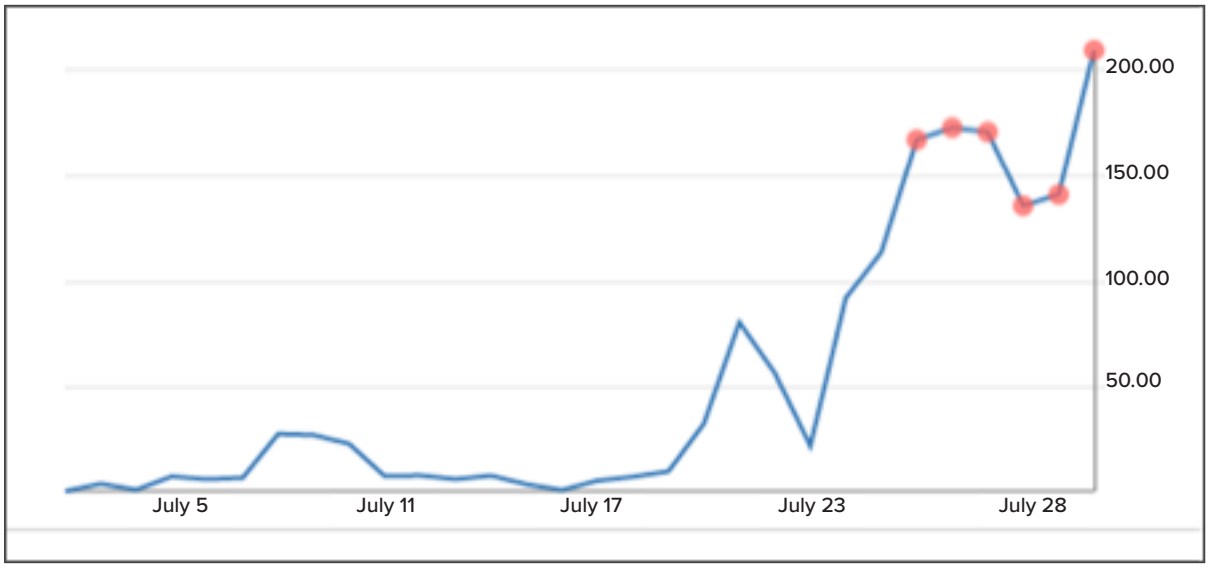
High-density leads to better utilization of resources. While Kubernetes and other orchestrators offer many functions, success derives from improving utilization of virtual machines. Why pay for 100% of a resource if you are only using 10%? Isn’t this the cost-reduction promise of the cloud? Once an established bin-packing algorithm is in use, an orchestrator can allow you to run more applications with fewer resources and lower costs.



Anomaly Detection

Organizations and tools are getting good at tracking top consumers and taking action on them. But an overlooked source of excess expenditure is what we like to call "**The Low-Grade Fever Problem**". This is a service that doesn't show up as a top consumer, but is still rising incrementally or has a sudden code change. Although the service doesn't show up as a top consumer, cloud consumption by the service, doubles. These anomalies typically remain undetected and if enough occur, can result in significant cost overruns.

Recently, AIS used an anomaly detection algorithm to help a client detect such changes. The following diagram depicts a consumption anomaly that our tool helped catch.



Azure Insights showing the overnight growth of an application

COST-AWARE COMPUTING

One aspect of “Cloud-Oriented Development,” which both deserves and has a whitepaper of its own, is “Cost-Aware Computing.” This is an aspect of cloud computing that has not been given a great deal of exposure. In a highly mature organization, one aspect of design and architecture is to consider the cloud consumption costs upfront. The following is an example of cost aware design: Azure Batch offers low-priority VMs (available at 80% discounts as surplus inventory). To take advantage of low-priority VMs, an application must be designed in a manner that is resilient to the availability of VMs.

Finally, a true DevOps mindset is needed to reduce operational costs. Developers (not just the operations folks) should be asking themselves how they can contribute to operational efficiency – whether it is developing a better metric for auto-scaling, or tracking idled resources.

CONCLUSION

Migrating to the cloud is the first step in an organization’s cloud journey. Much has been said about Cloud Maturity in the realms of tooling and governance, but the idea of Cost Maturity is only beginning to evolve.

This whitepaper focuses solely on a few of the techniques and concepts behind Cost Maturity Models, but it’s the start of a conversation that spans systems engineers, accountants, and developers. In the world of the cloud, just like security, cost management is everybody’s job now.

AIS CLOUD CREDENTIALS

AIS has been cutting our teeth with public cloud infrastructure since 2008 through our partnerships with Microsoft and AWS. Given our matchless Azure expertise, Microsoft turns to us to build Azure reference architectures and blueprints to migrate complex application environments to Azure. *And we've created a proven Cloud Adoption Framework* that provides step-by-step guidance and best practices to move to the cloud.

As a company built on a foundation of *application development* and software consulting capabilities, AIS is uniquely qualified to take our clients through all phases of cloud adoption. We can move clients from quick "lifting and shifting" of existing applications into cloud infrastructure services (IaaS) to refactoring applications harnessing native cloud platform as a service (PaaS) and software as a service (SaaS) capabilities.

We have helped scores of complex commercial enterprises and government organizations migrate to Azure, AWS, Office 365, and Dynamics 365. We will spend the time to understand your business needs and workload requirements and then outline a meticulous cloud plan that meets your scalability, governance, security, and budget needs.

GET STARTED WITH AIS TODAY

Contact AIS to begin your modernization journey. With the right people, expertise, and best practices in place, you can be sure you're on the right track to modernizing your apps.

Learn More at www.appliedis.com

ais

www.appliedis.com

11400 Commerce Park Drive, Suite 600, Reston, VA 20191
Phone: (703) 860-7800 Fax: (703) 860-7820

sales@appliedis.com