IT BUSINESS INTELLIGENCE

MANAGING CLOUD SERVER COSTS SUCCESSFULLY WITH ITBI™

Executive Summary

According to a recent international survey[1], 30 percent of all servers are completely idle or severely underutilized. Many installations attempting to address this issue have looked to cloud approaches such as Infrastructure as a Service (IaaS), where the customer can easily adjust capacity to the actual needs.

In reality, cloud solutions have often only made the problem worse. The driving force behind cloud is agility - additional cloud capacity can be acquired with the click of a mouse. The power to create servers and add capacity can be widely delegated within the organization, often outside the control of the IT organisation. This flexibility is generally used to create new servers or provide *under*-configured servers more capacity. Reducing the capacity of *over*-configured servers or stopping servers that get launched and forgotten, however, requires a concentrated effort based on an understanding of where there is excess or idle capacity.

In addition to standard data provided by the guest operating systems, cloud providers have APIs with a wealth of data about capacity and performance from their hypervisors and billing systems. But most customers struggle to create value from this data, and the cloud providers may not have a strong interest in helping reduce costs. The figure below shows some examples of what data is provided and what is really required.

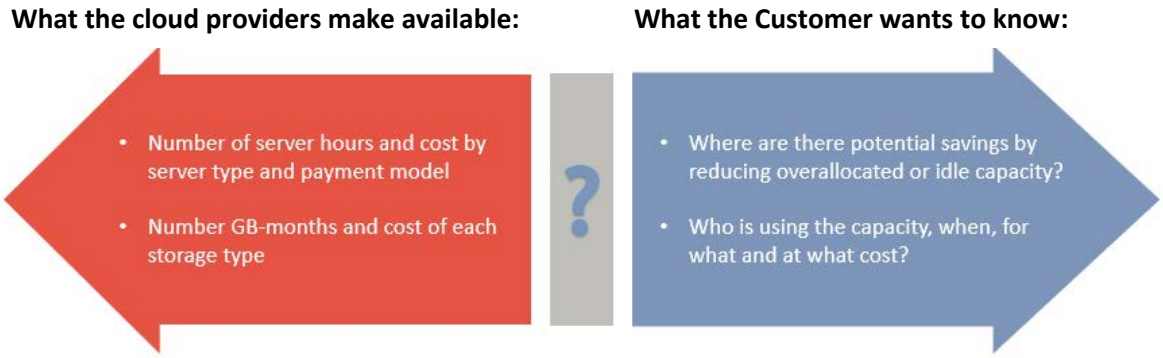

Figure 1: Examples of what cloud providers make available in terms of capacity and cost reporting versus what customers want to know

The answer to the first question, 'where are there potential savings by reducing overallocated or idle capacity', is the same for cloud as for in-house servers. The tools and methods to identify and rightsize underutilized servers, are discussed in Achieving Cost Savings by Rightsizing Servers with ITBI and are similarly applicable to the cloud environment.

This document focuses on answering '*Who is using the capacity, when, for what and at what cost*?' Answering that question requires a good data warehouse to manage the data and good reporting and analysis tools. It is also important to enrich the technical data with cost and organizational or application mapping in order to understand '*who is using what for how much'*. SMT Data's IT Business Intelligence (ITBI) solution is built for this and the examples in this document are based on SMT Data's experience using ITBI.

Managing costs in a cloud environment is an iterative process, getting you closer and closer to an optimal configuration and then proactively managing it on an ongoing basis going forward. This is especially true in a cloud environment where it is easy to extend the capacity. SMT Data, and our Business Partners, can provide both the tools and the consulting assistance required to ensure successful cost management.

With the right tools, organization and processes, and with a relatively small amount of time and effort, managing cloud server costs can save a large IT organization millions of dollars a year. The savings opportunities in most installations are significant and clear cut once the correct data is available.

The low hanging fruit can easily be identified with ITBI - even without knowing all the technical details or understanding the full complexity of the cost model.

[1] Research published in June, 2015 by Jonathan Koomey, Research Fellow, Steyer-Taylor Center for Energy Policy and Finance, Stanford University and Jon Taylor, Anthesis. www.koomey.com / www.anthesisgroup.com

Who is using what and for how much?

Cloud providers are very good at reporting how much capacity they have made available and billing for that capacity on a regular basis. The invoices are often very technical. The figure below, shows an example of part of an invoice from a cloud provider, specifying how many hours of each server and storage type was used and what it cost[2].

USD 0.226 hourly fee per Windows with SQL Server Standard, m3.large instance	4,320 Hrs	$976.32
USD 0.226 hourly fee per Windows with SQL Server Standard, m3.large instance	2,160 Hrs	$488.16
USD 0.45 hourly fee per Windows with SQL Server Standard, m3.xlarge instance	720 Hrs	$324.00
USD 0.45 hourly fee per Windows with SQL Server Standard, m3.xlarge instance	927.841 Hrs	$417.53
Windows with SQL Server Standard, m3.large reserved instance applied, m3.large instance used	4,320 Hrs	$0.00
Windows with SQL Server Standard, m3.large reserved instance applied, m3.large instance used	2,160 Hrs	$0.00
Windows with SQL Server Standard, m3.xlarge reserved instance applied, m3.xlarge instance used	720 Hrs	$0.00
Windows with SQL Server Standard, m3.xlarge reserved instance applied, m3.xlarge instance used	720 Hrs	$0.00
EBS		$5,371.42
$0.05 per GB-Month of snapshot data stored - EU (Ireland)	10,744.329 GB-Mo	$537.22
$0.055 per 1 million I/O requests - EU (Ireland)	671,047,935 IOs	$36.91
$0.055 per GB-month of Magnetic provisioned storage - EU (Ireland)	22,339.383 GB-Mo	$1,228.67
$0.11 per GB-month of General Purpose SSD (gp2) provisioned storage - EU (Ireland)	32,442.003 GB-Mo	$3,568.62

Figure 2: A section of an invoice from a cloud provider showing the usage and costs in technical terms

It is the responsibility of the customer to keep track of who is using the capacity and for what purpose. The cloud providers don't always make it easy to answer the following questions:

- Which servers are over-configured compared to the real capacity requirements?
- What are the potential savings from reducing the capacity of these servers or stopping them if they are idle?
- Whose servers are they – e.g. what business area or application?
- How much of the storage we are paying for is connected to servers that are currently stopped?
- How can the total cost be split out by business area or application so that the users are held accountable for the costs?

[2] For the sake of simplicity, the examples here focus on two of the most basic Infrastructure as a Service (IaaS) components – servers and storage. There are many other important models for delivering Cloud capacity and services, for example Platform as a Service (PaaS). There are also many other payment models than the simple allocated capacity based model described in this document - for example 'Reserved Instances' where the customer gets a discount for committing to purchase a certain amount of capacity for a certain period.

Rightsizing servers in a cloud environment

The tools and methods used to identify and rightsize underutilized servers for an on-site installation, also apply to the cloud environment. See Achieving Cost Savings by Rightsizing Servers with ITBI for a general discussion of server rightsizing.

Successful cost management in a cloud environment

Successful management of cloud costs requires tools that create transparency - combined with people and processes focused on cost-hunting. A good data warehouse to manage the data, and good reporting and analysis tools are a must. It is also important to enrich the technical data with cost and organizational information in order to understand *'who is using what for how much'*. SMT Data's IT Business Intelligence (ITBI) solution is built to solve that problem. The examples in this document are based on SMT Data's experience using ITBI.

1. Gathering and Enriching the Data

Data about capacity and utilization is fundamental to rightsizing. The data should be collected in a 'low touch' fashion. There are two general approaches to this in a cloud environment:

- Collecting the data from the APIs made available by the cloud provider. Amazon, for example, has a rich API (Amazon SDK) that ITBI uses to collect capacity, performance and cost data.
- Collecting the data from the virtual cloud servers themselves. ITBI uses the Windows Management Instrumentation (WMI) interface to remotely gather capacity and performance data from each cloud server. Similarly, standard Unix commands are executed through a secure shell connection to gather data from Linux and Unix servers. This data can be collected at the process or service level and is therefore often much more granular than the data from cloud provider's APIs.

The data from both sources should be integrated and stored in a well-structured data warehouse with good analysis and reporting tool support. ITBI is based on standard data warehouse and BI technologies and includes a rich set of standard reports to ensure a simple implementation and fast time-to-value.

The technical data can then be enriched with business information. Technical dimensions such as server name can be mapped to a business dimension such as the organizational unit owning the server or the application running on the server or preferably both. This mapping can often be derived from a naming or tagging convention used when the server was launched.

The technical measurements can also be enriched with cost information. Pricing models for cloud environments such as Amazon EC2 are publicly available (see for example http://aws.amazon.com/ec2/pricing).

Exact costing can be quite complicated, for example due to different payment models such as 'reserved instances' or 'spot instances'. But approximate costs are good enough for the purpose of cost optimization and rightsizing. The objective is to identify the most expensive servers with the lowest utilization in order to know what to focus on.

2. Understanding Costs in Technical Terms

In a typical cloud environment, the customer pays for a server based on its configuration (e.g. number of cores, amount of RAM, software included) and the number of hours that the server is running. New servers can easily be started and easily reconfigured with more (or less) capacity as needed. Customers generally only pay for the hours that the server is running. If you stop a server when you are not using it, then it doesn't cost anything (though, as discussed below, you will normally still pay for the storage attached to the server).

The graph below shows the daily cost for servers and storage growing over a period of time:

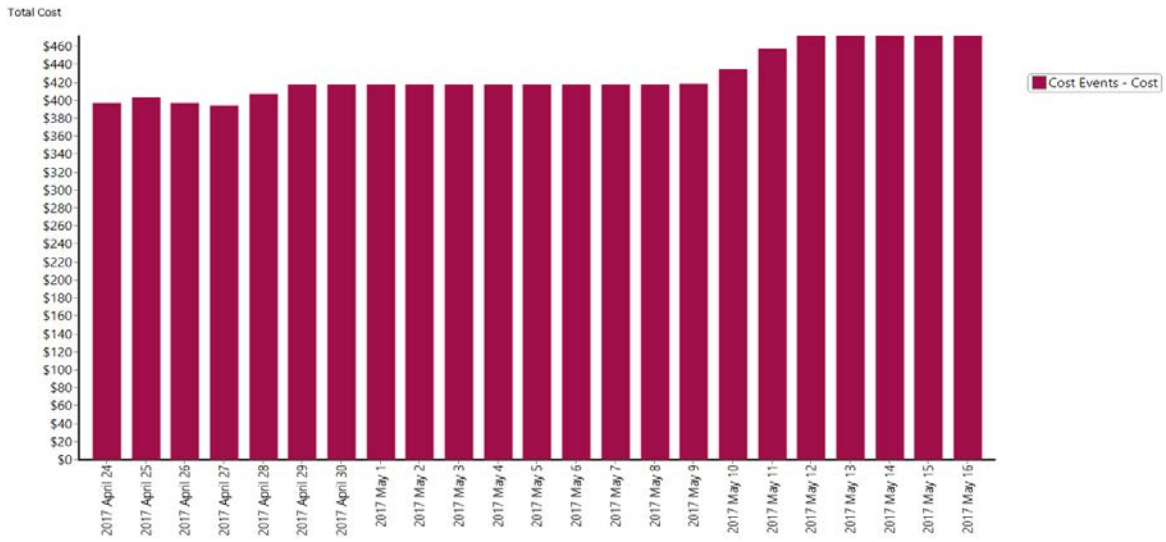

Figure 3: Total cost per day of servers and storage

To understand why the cost is growing, we can start by breaking down the cost into server cost and storage cost:

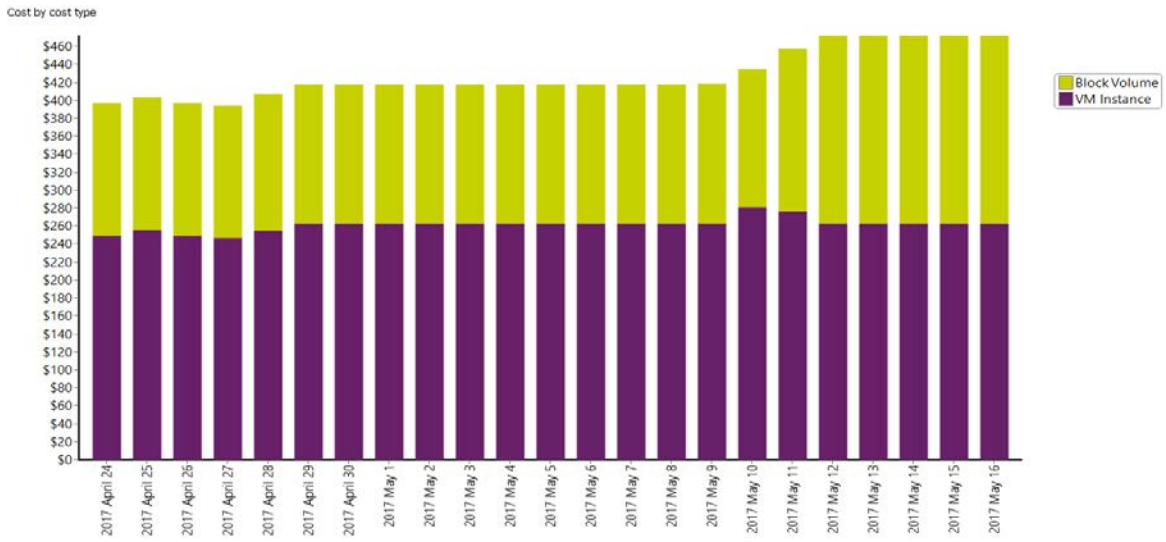

Figure 4: Break down of total cost into servers (VM Instances) and storage (Block Volume)

Here we see that the increase in cost April 28-29 is primarily due to more servers being started while the increase May10-12 is primarily due to additional storage.

We can further break the server cost down by server type:

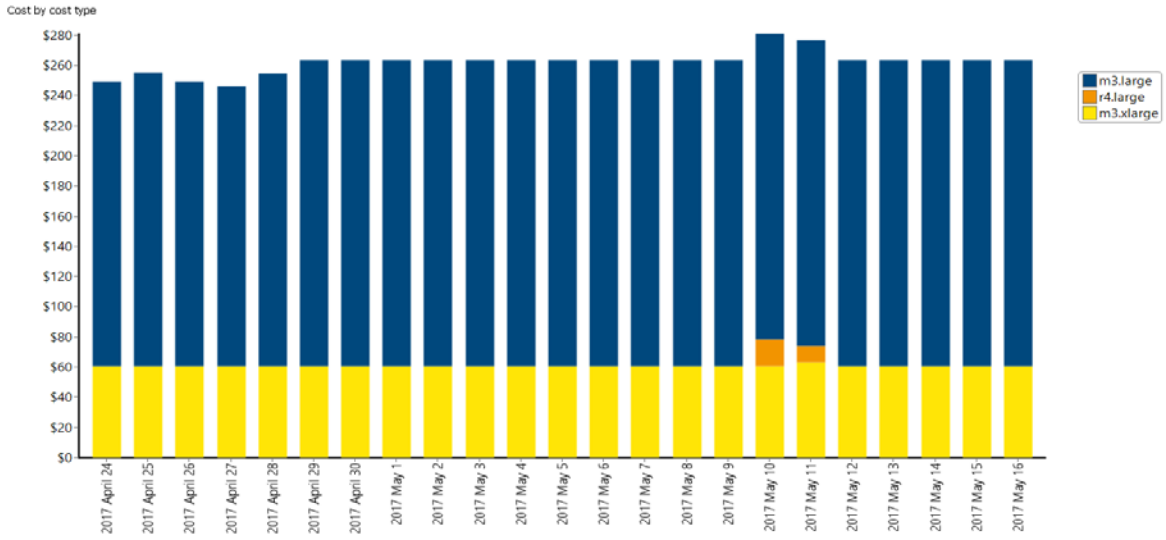

Figure 5: Further break down of server cost component by server type

Above the increase in April is due to additional servers of type m3.large (a server with 2 cores and 8GB of RAM), and that there was a temporary spike May 10-11 where an r4.large (2 cores and 15GB of RAM) server was in use.

Similarly, we can break the storage cost down by type:

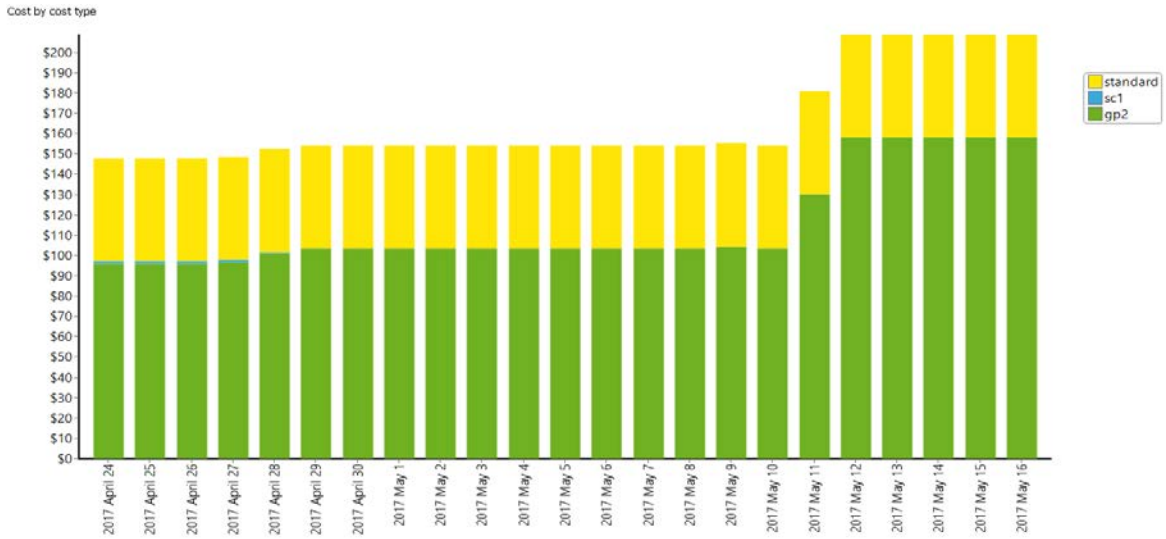

Figure 6: Further breakdown of storage cost by disk type

The figure shows that the increase in May is due to additional gp2 (general purpose SSD) being used.

We can also look at storage usage by state:

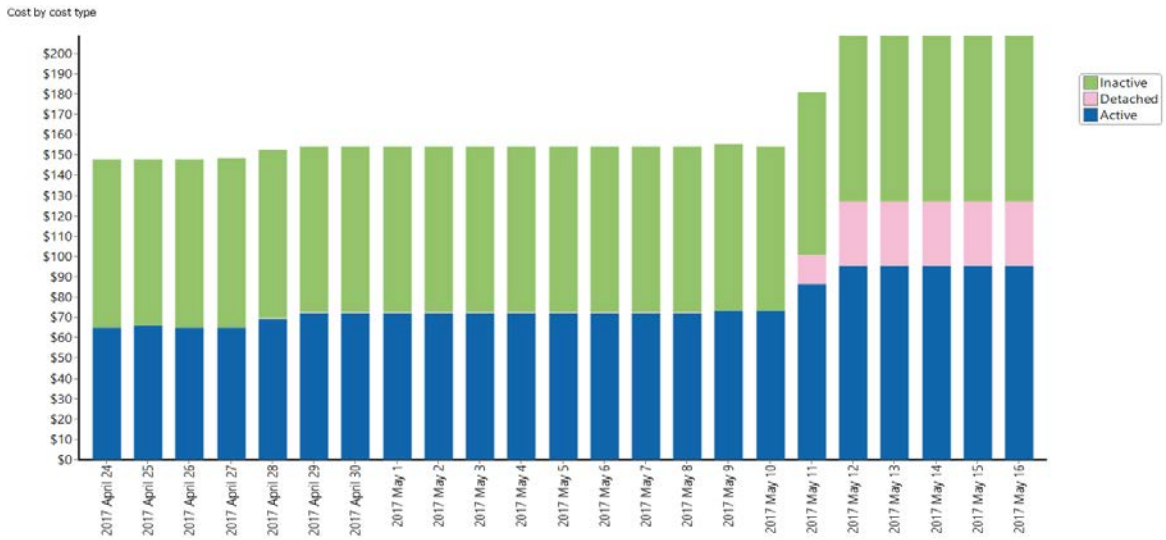

Figure 7: Further breakdown of storage cost by whether the storage is in use (Active), attached to a server that is turned off (Inactive), or no longer attached to any server (Detached)

The figure shows that a significant portion of the storage cost is inactive or detached. This is storage that is being paid for, but not currently in use. Inactive storage is attached to a server that is not currently running. Detached storage is no longer attached to a server at all. We can see that the increase in storage cost in May, is both due to additional active storage but also to an increase in detached storage.

Customers often learn to stop servers when they are not in use, but are often surprised by the fact that their cost of storage continues to grow. Initially this cost is quite small because the cost of storage is small relative to the cost of the server. But over time as new servers are started and the old servers are stopped, the amount of storage that is attached to stopped servers continues growing. All because the decision to stop a server is easy to make – you can always start it again if it is needed. The decision to terminate storage (and delete the data or move it to a cheaper medium) is a more complex decision and therefore often one that gets postponed.

3. Understanding Costs in Business Terms

Cloud providers typically report the number of server hours of each type of server and GB-months of each storage type and what that costs. The reporting they provide is at a technical level.

What's needed though is not just a technical understanding of capacity usage but also insight into who is using the capacity and for what purpose. ITBI enriches the technical data with business information, describing for example which business area, application, or environment is responsible for the capacity usage, and what function is being carried out.

For example, if we want to understand the usage shown in the section above in business terms, we can map the costs to the business areas that are responsible for the servers in question. There are many ways of doing this. One simple approach with Amazon EC2 is to tag the servers with a business area name when the servers are created. ITBI can then extract the tagging information along with the capacity and consumption information.

The example below, pictures the same cost curve we analysed in the previous section, but now broken down by who is using the capacity based on a mapping of servers to business areas. In a similar way, we could have mapped the costs to applications, environments or other business dimensions.

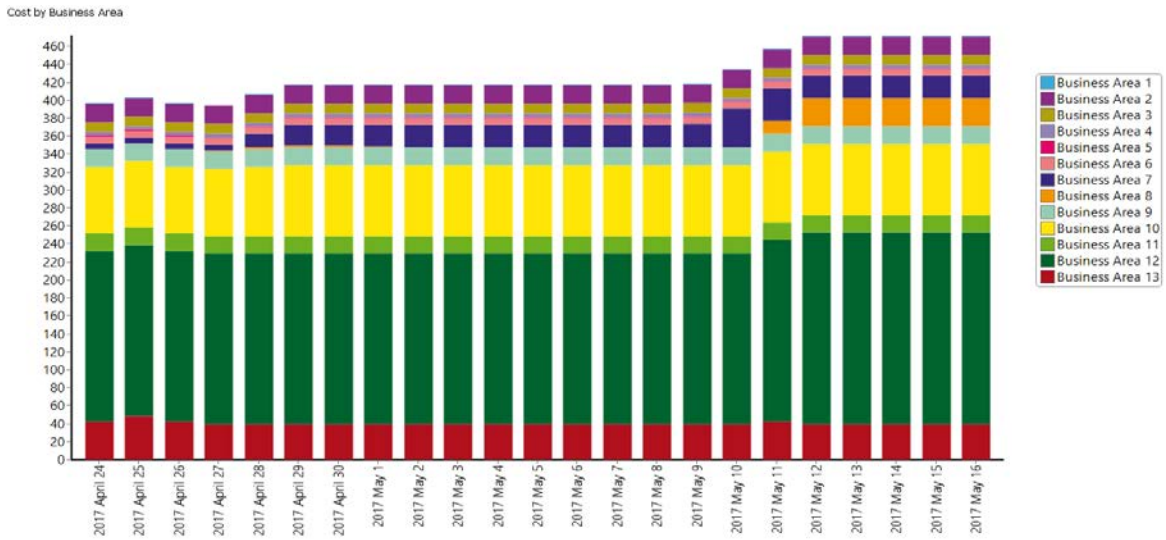

Figure 8: Total costs allocated to who is using the capacity based on a mapping from server (and attached storage) to business area

The figure shows that the increase in cost April 28-29 is due to Business Area 7 and the increase May 11-12 is due to Business Areas 8 and 12.

ITBI allows the business dimensions to be easily combined with the technical ones to drill down and further understand cost drivers.

For example, we can analyse the cost for Business Area 7 and split it out by type:

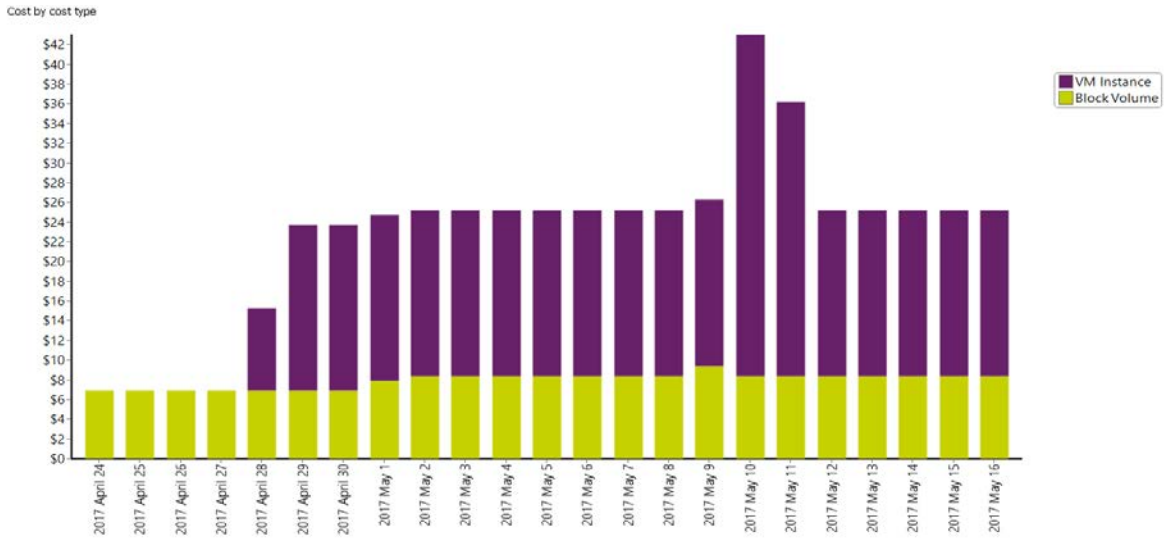

Figure 9: Cost breakdown for a selected business area (Business area 7) into server cost and storage cost

Above we see that the cost for this business area was just storage until April 28, when servers in this business area were started up. Additional servers were also temporarily started for this business area May 10-11. Again, note that the storage cost remains more or less constant, but we only pay for the servers while they are actually running.

We can also drill down and look at the storage cost for this business area:

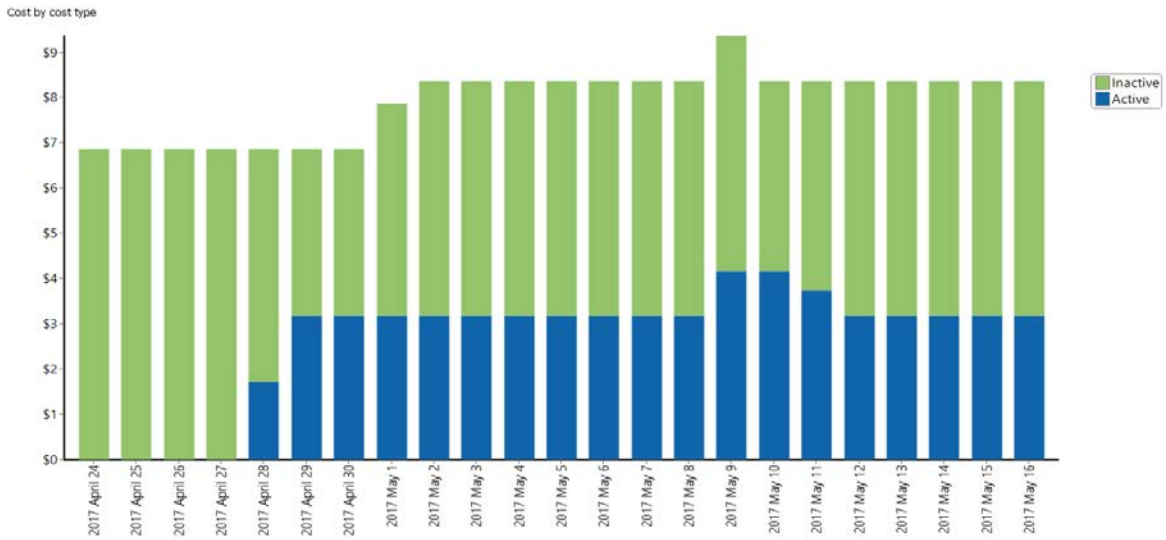

Figure 10: Storage cost breakdown for a selected business area (Business Area 7) into active and inactive disk

The figure above shows us that the only cost for this business area was 'inactive' disk until April 28. At this point in time, as we saw above, one or more of the inactive servers was started up and the disk became 'active' again.

We can continue to drill down using standard ITBI functionality to the individual disks, the individual servers and even the individual processes and services running on these servers. This can help us understand not only our allocated capacity, but also if we are using that allocated capacity effectively.

For more on this see Achieving Cost Savings by Rightsizing Servers with ITBI for a general discussion of server rightsizing.

Conclusion

Many large IT installations are well aware that better capacity management including rightsizing servers can lead to huge cost savings. In many cases the desire to be more in control of the allocated capacity, and just pay for what is needed, has driven the move to the cloud. Unfortunately, cloud has, in many cases, just exasperated the problem. Customers are overwhelmed by the perceived complexity in understanding and managing cloud capacity and costs. They lack good data or tools to create transparency into the ever-growing jungle of servers and complicated cost models.

With the right tools, organization and processes, and with a relatively small amount of time and effort, a large IT organization can save millions of dollars a year by managing cloud costs more effectively. The savings opportunities in most installations are large and clear cut.

The low hanging fruit can be easily identified with ITBI even without knowing all the technical details or understanding the full complexity of the cost model. SMT Data, and SMT Data's partners, can provide both the tools and the consulting assistance required to ensure cost management efforts are successful.

About SMT Data and IT Business Intelligence

SMT Data, a software and services company based in Denmark, has developed a unique software solution that collects, aggregates, and processes enormous amounts of technical data from the company's IT-infrastructure. Conceptually it is Business Intelligence for IT – we call it ITBI.

SMT Data delivers services, knowledge and tools to help you manage your data centre more efficiently, monitor your outsourcing and Cloud providers and link the use of IT directly to the company's bottom line and development potential. For 27 years, we have supplied fact-based optimization:

- *Optimize IT infrastructure utilization and performance*
- *Link IT resource consumption and IT costs to business activity*
- *Control outsourcing providers (including cloud providers) and optimize outsourcing costs*
- *Connect application development to IT operations – DevOps*
- *Consolidate assets and balance load (for M&A, virtualization etc.)*
- *Reduce time spent on analysing and reporting*

The collection, selection and transformation of the technical data is governed by an extensive set of rules and policies that embody SMT Data's deep knowledge of how this data can be used and understood. A fully automated process transforms terabytes of unstructured technical data to just gigabytes of information in a well-structured data warehouse running either in the cloud (ITBI as a Service) or at your datacentre.

The data can be easily analysed either using the extensive set of ITBI standard reports/analysis or by developing new reports using the standard end-user BI tool. The technical reporting works 'out of the box' and creates immediate value by identifying capacity, performance and demand optimization potential.

Steven Thomas, CTO, SMT Data

October, 2017

Website: www.smtdata.com
Contact: info@smtdata.com