Capacity planning in marketing

Received (in revised form): 30th December, 2020

Andrew Pearson
Managing Director, Intelligencia, Hong Kong

Andrew Pearson is the Managing Director of Intelligencia Limited, a leading implementer of artificial intelligence, business intelligence, data warehousing, data modelling, predictive analytics, data visualisation, digital marketing, mobile, social media and cloud solutions for the gaming, finance, telco, hospitality and retail industries. He has a degree in psychology from UCLA, and has worked in such sectors as IT, marketing, mobile technology, social media and entertainment. He writes on a variety of topics, including mobile media, social media, predictive analytics and cloud technology, and his work has appeared in numerous magazines and journals.

Intelligencia, 505 Hennessy Road, Suite 613, Causeway Bay, Hong Kong
Tel: +852 5196 1277, +853 6616 1033; E-mail: andrew.pearson@intelligencia.co

Abstract Capacity management is the process of optimising production in line with fluctuating demand for products and services in order to reduce wasted capacity. Simply put, it aims to make optimal use of essential resources while minimising the use of non-essential resources. To this end, the key is to balance the right number of users and the right performance at peak usage to ensure a great end-user experience. This paper explores a capacity planning solution that can predict upcoming costs with advanced predictive analytics and forward-thinking what-if scenario modelling that can produce a healthy return on investment as well as help companies go green.

KEYWORDS: capacity planning, real-time monitoring, AIOps, hyperautomation, application demand modelling, cloud bursting, predictive resource scaling, predictive analytics

KNOW YOUR LIMITATIONS

> ‘A constraint limits or holds back the possible success of an operational strategy. The theory of constraints, an organisational change method focused on process improvement, contends that every organisation must face at least one constraint, or weak link in the process chain. In regard to marketing, constraints may affect product, price, place or promotion.’ (John DuBois)[1]

The theory of constraints (TOC), conceived by Dr Eliyahu Goldratt,[2] is a highly focused methodology for prioritising improvement activities in order to generate rapid gains. The methodology offers the following benefits:

- fast improvement;
- greater capacity;
- increased profit;
- reduced lead times; and
- reduced inventory.

The TOC is well suited to marketing systems as it can help Chief marketing officers (CMOs) understand their departments, their systems and their data in a way that will help them make better and more profitable decisions. Whereas the primary focus of the TOC is the rapid improvement of constrained processes in order to realise more profit, it is a waste of time optimising non-constrained processes and systems as they will not produce any significant benefits, financial or otherwise.

As Ira Kalb claims, ‘the primary way a CMO can prove his or her worth is to collect the data on the return the company is realising on its marketing investment.

To do that, a comprehensive marketing information system is required'.[3] In a perfect world, Kalb argues, businesses would be able to integrate their marketing information system (MIS) into their various systems and processes. In this rather idealistic world, every sale, lead and marketing offer would be traced back to the marketing effort that produced it. Every dollar of marketing spend could be valued and quantified. Every compliment or complaint could be tracked back to its source as well, bringing new measurable value to social media. This is a worthy endeavour, of course, but not a simple one.

'Rather than wait for the dream to materialise, marketers need to improvise. They need a system that enables them to (1) make better decisions and (2) support those decisions with verifiable data', says Kalb.[3] Systems should be checked and itemised, with marketing information data gleaned from them. A gap analysis should be initiated to identify what information is not being provided to the marketers in the current system. Additional systems that can provide the needed marketing information should also be created. To the extent possible, says Kalb, these systems should be integrated company-wide.

'We are drowning in information but starved for knowledge', said John Naisbitt, some 40 years ago.[4] Since then, this idea has only grown in relevance. Of the seven 'V's of Big Data — volume, velocity, variety, variability, veracity, visualisation and value — volume is growing exponentially. Data collection, correlation and data use are not only accelerating massively but increasingly doing so in real time; indeed, this is becoming a minimum requirement for many IT departments because their customers require nothing less.

With the ability to collect, append, as well as send to or receive data from almost anywhere in the world, data variety and variability are increasing substantially. The veracity or accuracy of data is also getting easier to ensure as many data integration tools have data cleansing and data verification capabilities built in. The visualisation of data has become exceptionally easy thanks to business intelligence tools like Domo, PowerBI, Qlik and Tableau, which have simplified the creation of dashboards. The growing value of data is unquestionable — out of the ten most valuable companies in the USA, five depend on data for their lifeblood.

Credit card companies collect every penny we spend. Social media companies gather every like and dislike we make. Internet of things (IoT) devices capture sensor data of all kinds — and there will soon be an explosion of such devices. Gyroscopes and accelerometers collect every physical movement of our phones, while mobile apps collect highly valuable personal information, including our every location. Never before have so few tracked so many, while also profiting so handsomely from so much captured data. And yet, while there have never been so many ways to collect, track, quantify, metatag and visualise data, many marketers remain ignorant when it comes to using this information effectively.

According to the 'Flexera 2020 State of the Cloud Report',[5] organisations are wasting 30 per cent of their cloud spend, paying too much for services they simply do not need. This translates to a huge sum of money being flushed away, to say nothing of the squandered energy. Indeed, going green is a competitive advantage today, so this is an opportunity going to waste, literally. According to the report:

> 'some of the increase is a result of the extra capacity needed for current cloud-based applications to meet increased demand as online usage grows … Other organisations may accelerate migration from data centres to cloud in response to reduced headcount, difficulties in accessing data centre facilities and delays in hardware supply chains'.[5]

Philip Kotler defines a marketing information system as a 'continuing and interacting structure of people, equipment and procedures to gather, sort, analyse, evaluate, and distribute pertinent, timely and accurate information for use by marketing decision makers to improve their marketing planning, implementation, and control'.[6] One of the tools that can help with control is capacity planning, which aims to minimise the discrepancy between the capacity of an organisation and the demands of its customers. As Zoltán Sebestyén and Viktor Juhász claim, 'capacity is one of the most important measures of resources used in production. Its definition and analysis are therefore one of the key areas of production management'.[7]

Capacity demand varies according to changes in production output, such as increasing or decreasing the production of an existing product or producing new products. Better utilisation of existing capacity can be accomplished through improvements to the effectiveness of equipment. Capacity can be increased by introducing new techniques, equipment and materials, increasing the number of workers or machines, increasing the number of labour shifts, or acquiring additional production facilities. In a nutshell, capacity planning aims to make the best possible use of essential resources while minimising the use of non-essential resources.

Capacity planning works hand-in-hand with Kalb's condition that every dollar spent on marketing be valued and quantified.[3] It attempts to ensure there is as little wastage as possible when it comes to the IT estate, including with the marketing department's IT activities. The question is, can marketing's IT activities be quantified so that every sale, lead, marketing offer, and the cost of the employees and software handling these activities be traced back to the marketing effort that produced them? This would help enormously with justifying return on investment (ROI).

According to Erik Brynjolfsson:

> 'The critical question facing IT managers today is not "Does IT pay off?" but rather, "How can we best use computers?" … Even when their IT intensity is identical, some companies have only a fraction of the productivity of their competitors'.

Unlike a certificate of deposit, an investment in IT does not produce an expected or guaranteed rate of return.[8] An IT estate has countless moving parts — both virtual and physical — from the software running atop it, to the people overseeing it. As the innumerable variables interact with each other, weathering unintended demand spikes, calculating a definitive ROI for many IT-related costs becomes almost impossible.

Furthermore, contends Brynjolfsson:

> 'IT is only the tip of a much larger iceberg of complementary investments that are the real drivers of productivity growth … In fact, our research found that for every dollar of IT hardware capital that a company owns, there are up to $9 of IT-related intangible assets, such as human capital — the capitalised value of training — and organisational capital — the capitalised value of investments in new business-process and other organisational practices. Not only do companies spend far more on these investments than on computers themselves, but investors also attach a larger value to them'.[8]

Brynjolfsson cautions:

> 'Too often, the flow of information speeds up dramatically in highly automated parts of the value chain only to hit logjams elsewhere, particularly where humans must get involved and processes are not updated. The result is little or no change in overall performance. A gigabit Ethernet network does no good if the real bottleneck is a manager's ability to read and act on the information … In the information economy, the scarce resource is not

> information, but the capacity of humans to process that information'. [8]

MONITORING AND MEASURING EVERYTHING

Man's role as 'the best condition monitoring device ever invented'[9] is under threat from such tools as business process management software, robotic process automation, hyperautomation, artificial intelligence operations (AIOps) and real-time monitoring. While the experience of an IT technician who intimately understands every nuance of a system he has been working with for years is priceless, today's monitoring and alerting systems can go further as they can help systems to self-heal.

Gartner defines AIOps as a platform that utilises:

> 'Big Data, modern machine learning and other advanced analytics technologies to, directly and indirectly, enhance IT operations (monitoring, automation and service desk) functions with proactive, personal and dynamic insight. AIOps platforms enable the concurrent use of multiple data sources, data collection methods, analytical (real-time and deep) technologies, and presentation technologies'.[10]

AIOps analyses a system's data to learn about a company's day-to-day operations, including marketing, then either fixes the issues it finds or proactively attempts to fix the potential issues it sees coming down the pipe.

According to Trent Fitz:

> 'modern apps are comprised of millions of containers and serverless functions strewn across multiple clouds, and each one of these application components may exist for days or less than a second. Stitching all of this information together while trying to find outliers is magnitudes more difficult than trying to isolate a rogue Java thread on a typical application server'.[11]

It is an enormous task to monitor memory usage and CPU utilisation while overseeing the spin-up and spin-down of clusters, observing server fan speeds and ensuring error alerts need checking to ensure they are not superfluous. Figure 1 shows the monitoring setup of a typical data centre.

But help is on its way. Atul Soneja explains the AIOps process when alerts occur:

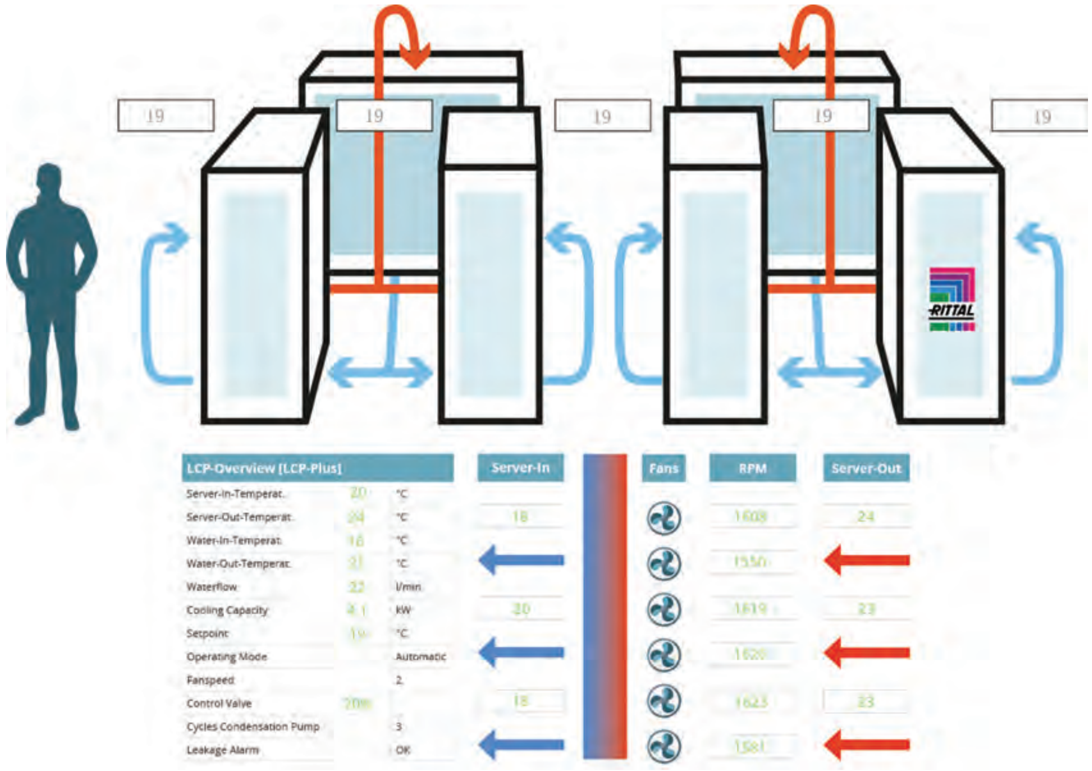
> 'The AIOps solution automatically opens the ticket and enriches it with log information, events, and metrics before directing it to the right person. Now, all the information is already there, and IT knows what to do with it. All of this is handled automatically behind the scenes, so teams never have to close a ticket manually again'.[12]

Today's AIOps solutions collect metrics and logs, collate event streams with dependency data, and deliver end-to-end three-dimensional windows into a company's operations system. As Trent Fitz sees it, 'this means eliminating the no. 1 problem AIOps tools have experienced thus far, ie limited visibility and context due to the lack of cardinality in the data they're analysing'.[11] This not only enhances the system's ability to understand a problem and initiate a solution but it also provides better visibility into the entire operation in an activity-based costing way. Granular details on the data and how the organisation uses this data becomes available and this information can be collated and quantified and made available to departments throughout the company.

CLOUDBURSTING

According to Sebestyén and Juhász:

> 'There are three aspects of the problem of conventional capacity measures: the absence of economic content, quantity based approach, and the unduly high emphasis laid on technical processes …



LCP-Overview [LCP-Plus]		
Server-In-Temperat.	20	°C
Server-Out-Temperat.	24	°C
Water-In-Temperat.	16	°C
Water-Out-Temperat.	21	°C
Waterflow	22	l/min
Cooling Capacity	4.1	kW
Setpoint	19	°C
Operating Mode		Automatic
Fanspeed		2
Control Valve	20%	
Cycles Condensation Pump		3
Leakage Alarm		OK

Figure 1: Monitoring a data canter
Source: ITRS.

> If capacity measures could side-step the problems discussed above, ie if they could include the value of resources, and could refer to the costs of unused capacity, then better decisions could be made in a number of cases'.[7]

As Sebestyén and Juhász argue, 'changes in the nature of production, and the enhanced significance of auxiliary processes made calculations necessary for production and service systems where processes are difficult to quantify'.

Adding a cloud component to a company's IT department is almost a given these days, but services like AWS, Azure, Google Cloud, Alicloud, and other cloud providers can be pricey. Managed cloud services are even worse. As Lee notes:

> 'Capacity planning is a challenging task when there is an unpredictable, fluctuating computing demand with many peaks and troughs. Without a solid evaluation model, estimating the tradeoff between the benefits and costs incurred in order to cover peak computing demand is challenging. Therefore, overcapacity or under-capacity is a common phenomenon in the investment of cloud capacity. Overcapacity puts companies at a cost disadvantage due to a low utilisation of cloud resources. On the other hand, under-capacity puts them at a strategic disadvantage due to customer/user dissatisfaction, high penalty costs, and potential sales loss'.[13]

Araujo *et al.* argue that 'the efficient and accurate assessment of cloud-based infrastructure is essential for guaranteeing both business continuity and uninterrupted services for computing jobs'.[14] However, López-Pires and Barán suggest that this is easier said than done as the efficient resource management of cloud infrastructures is highly challenging.[15] For Lee, most capacity planning and management area studies focus on micro-level scheduling such as dynamic resource allocation and prioritisation of computing jobs.[13] Widely used resource management methods like AWS's Auto Scaling and Azure's Autoscaling Application Block are both reactive; they are also provided by companies in a fox-guarding-the-henhouse kind of way, which is to say, their profits are based on usage, so are they really the best companies to provide applications that recommend use limitations?[13]

Balaji *et al.* suggest that 'while these reactive approaches are an effective way to improve system availability, optimise costs, and reduce latency, it exhibits a mismatch between resource demands and service provisions, which could lead to under or over provisioning'.[16] Several authors, including Wang *et al.*,[17] Han *et al.*[18] and Deng *et al.*[19] recommend predictive resource scaling approaches that can overcome the limitations of this reactive approach.

Laatikainen *et al.*[20] believe the hybrid cloud can reduce the financial burden of overcapacity investment and technological risks related to the full ownership of computing resources as well as allow companies to operate at a cost-optimal scale and scope under demand uncertainty.

With a hybrid cloud, companies can scale their computing requirements beyond the private cloud and into the public cloud — a capability also known as cloud bursting.[21] An application runs in its own private resources for most of its computing needs and then bursts into a public cloud when its private resources are unable to cope with surges in computing demand. For example, a popular and cost-effective way to deal with the temporary computational demand of Big Data analytics is a hybrid cloud bursting solution that leases temporary off-premise cloud resources to boost overall capacity during peak utilisation.[22]

While potential benefits of the hybrid cloud arise in the presence of variable demand for many real-world computing workloads, additional costs related to hybrid cloud management, data transfer and development complexity must be considered.[23] Everything from bandwidth, latency, location of data, and communication performance must be considered when integrating a public cloud with a private cloud.[24]

As each cloud provider has its own propriety system, there are no standardised solutions challenges, hence cloud users must integrate diverse cloud services obtained from multiple cloud providers and then perform cloud bursting in the hybrid cloud environment. While various standardised solutions have been developed for diverse cloud computing services, cloud providers often develop their own proprietary services as a way to lock in clients, differentiate their services, and achieve a market monopoly in the early stages of innovation.[25]

According to Forrester,[26] in 2018, cloud computing became a must-have technology for every enterprise. According to Lee, 'nearly 60 per cent of North American enterprises are using some type of public cloud platform. Furthermore, private clouds are also growing fast, as companies not only move workloads to the public cloud but also develop on-premises private cloud in their own data centers'.[13] Lee argues the corporate adoption of hybrid cloud computing is an irreversible trend because the demand for Big Data, smartphones and IoT technologies will not be receding any time soon.[13]

CAPACITY MANAGEMENT

Capacity planning attempts to reduce constraints within the IT estate. The TOC 'seeks to provide a precise and sustained focus on improving the current constraint until it no longer limits throughput, at which point

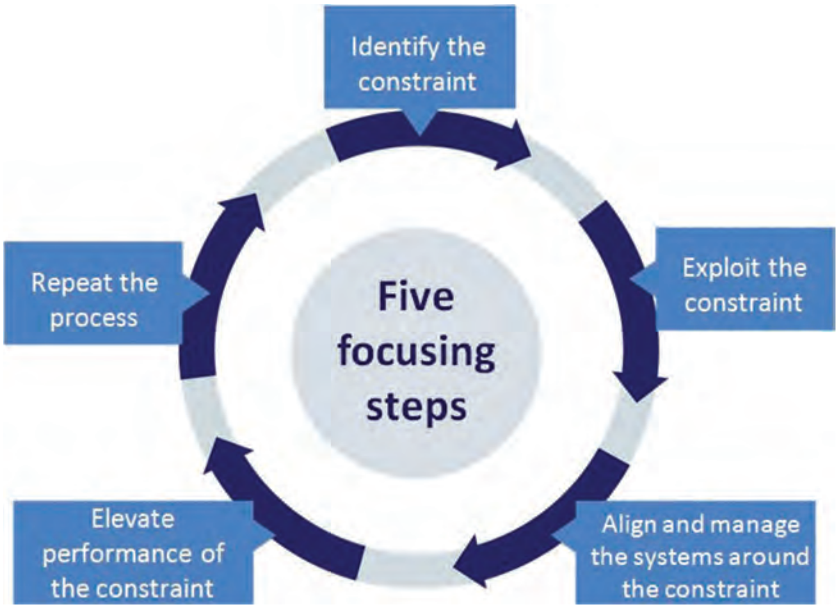

Identify the constraint
Exploit the constraint
Align and manage the systems around the constraint
Elevate performance of the constraint
Repeat the process
Five focusing steps

Figure 2: The theory of constraints uses a process known as the five focusing steps to identify and eliminate constraints (ie bottlenecks)
Source: Stefanova, H. (2014) 'Five Focusing Steps', available at: https://blogs.3ds.com/delmia/uncovering-hidden-lessons-goal-part-1/ (accessed 30th December, 2020).

the focus moves to the next constraint'.[27] The TOC lays out a five-step process known as the five focusing steps for identifying and eliminating constraints (see Figure 2).

The key to capacity management is counterbalancing the right number of users with the right performance at peak usage to ensure a great end-user experience. In the words of VMWare, the cloud services provider, 'Demand drives stress … Because the demand for capacity fluctuates in each environment, the top contenders for priority often include high efficiency versus low risk of poor performance'.[28]

'The stress concept involves how high and how long the demand persists relative to the capacity available', explains VMWare. This value is used 'to measure the potential for performance problems. The higher the stress score, the worse the potential is for degraded performance'.

Efficiency and optimisation are the goals for capacity planning. In an ITRS use case that looked at application demand modelling on a mobile banking platform, the client had the following three objectives:

- predict maximum call volume on the current architecture and identify degradation in performance as volumes increase;
- recommend changes to improve capacity limits and reduce unused infrastructure; and
- allow modelling of increase in volumes and predict impact.

For this use case, which could easily be extrapolated to a marketing department, the infrastructure data were provided by vCentre to develop the baseline, with a total of 1,420 virtual machines and 30 hosts. AppDynamics data were used to identify the servers associated with the mobile application and provided 'in-app' data. The metrics were as follows:

- business volume metrics — 'calls per minute';

- performance metrics — transaction response times;
- in-app metrics — time per minute spent on garbage collection; and
- tier/role of each server.

The customer wanted a view of performance as well as resource utilisation, which included the following standards:

- Over 90 per cent of transactions going through the system fell into one of the same four groups:
  - login;
  - check security question answer;
  - get balance; and
  - check transaction history.
- It was established that the 95th percentile of transaction responses would be less than three seconds. Anything over that would be considered a drop in performance and noticed by end users.
- For the Jboss servers, specific attention was paid to the time spent on garbage collection, ie attempting to reclaim memory occupied by objects that are no longer in use by the program. The standard set: this should be less than 10 per cent and the threshold for this metric was set at six seconds (10 per cent of a minute).

In general, ITRS recommends expressing capacity in business terms by understanding volume constraints and trends. For a deep dive into relationships, ITRS recommends the following steps:

- detect relationships between volume drivers and resource utilisation;
- identify the strongest relationships that drive application constraints;
- track ongoing behaviour and changes;
- create statistical summaries at multiple levels of the organisation and build up an accurate understanding of application behaviour and metric patterns; and
- normalise data from the multiple tools and technologies a company has throughout its organisation.

Insights from these activities should provide information about resource utilisation, implementation metrics, business volumes and critical transaction response times. The goal is to have a single, comprehensive view of the company's entire systems and all its processes.

ITRS's Capacity Planner solution (see Figure 3) attempts to understand the correlations between business and performance metrics. By understanding a system's relationships, ITRS's Capacity Planner determines the primary constraints in the application, including ones for marketing.

The tier constraint chart (see Figure 4) shows the maximum volume that each component of the application can handle before running into capacity limit issues. VM 1001 is the Jboss server with the lowest capacity for volumes, causing garbage collection issues once calls per minute volume reaches 4,314.

Using these mined relationships, a 'demand model template' is created so these can be used in a forward-thinking model. This allows the user to model an increase of up to 6,000 extra calls per minute (see Figure 5). At just over an extra 4,000 calls per minute, it is clear that the Jboss servers will breach their thresholds. A strong correlation between calls per minute and garbage collection rates (0.97 r-squared) is identified.

Figure 6 shows that, by the last operation, having added an extra 6,000 calls per minute, the issue becomes application-wide. In this case, it was predicted that 5 per cent of login response time would be above three seconds.

The recommendation is to drag and drop four new Jboss Virtual Machines (VMs) and distribute the demand evenly. Green icons on the timeline above indicate that

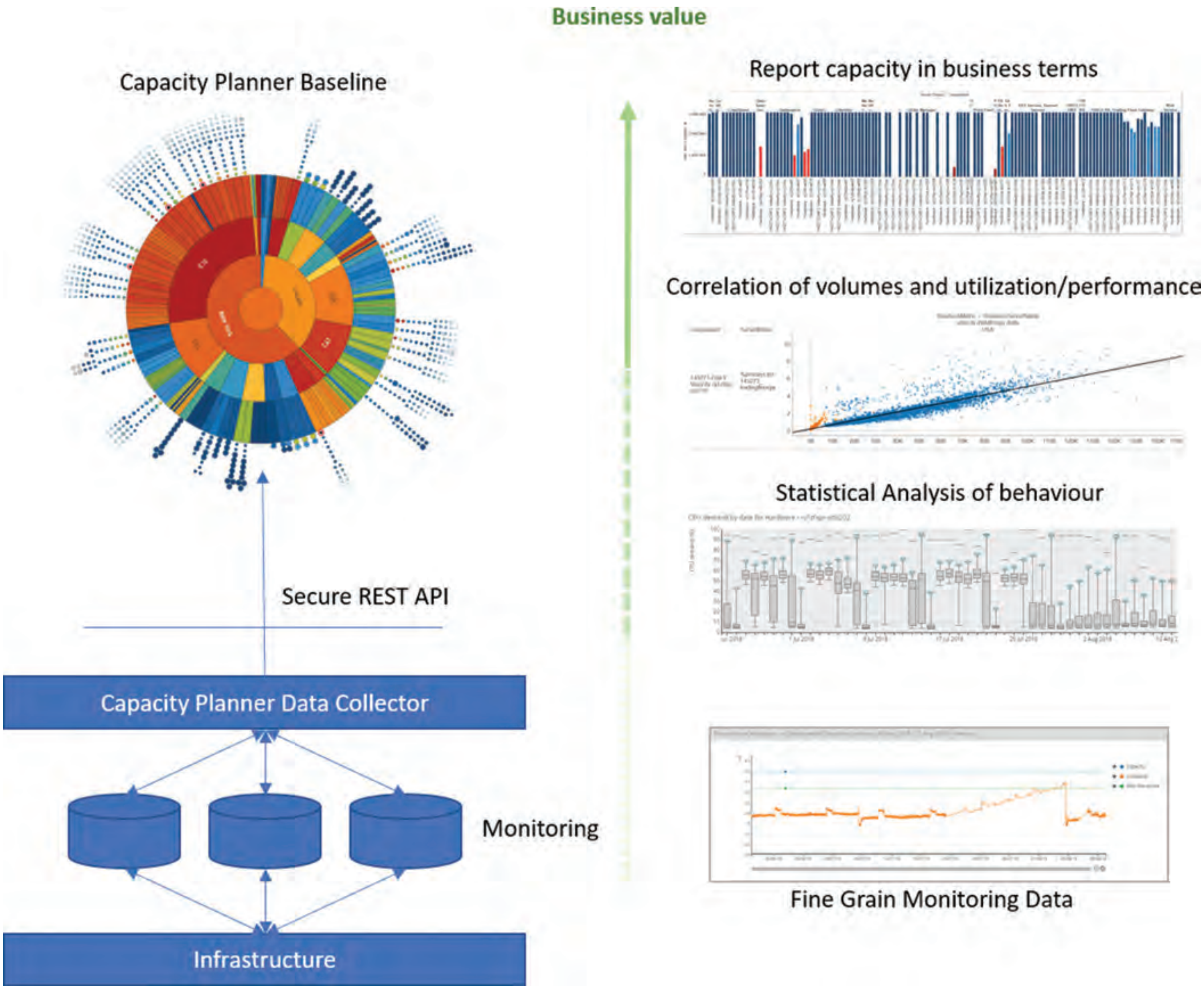

Business value
Capacity Planner Baseline
Report capacity in business terms
Correlation of volumes and utilization/performance
Statistical Analysis of behaviour
Secure REST API
Capacity Planner Data Collector
Monitoring
Fine Grain Monitoring Data
Infrastructure

Figure 3: The Capacity Planner architecture, showing how business value is derived
Source: ITRS.

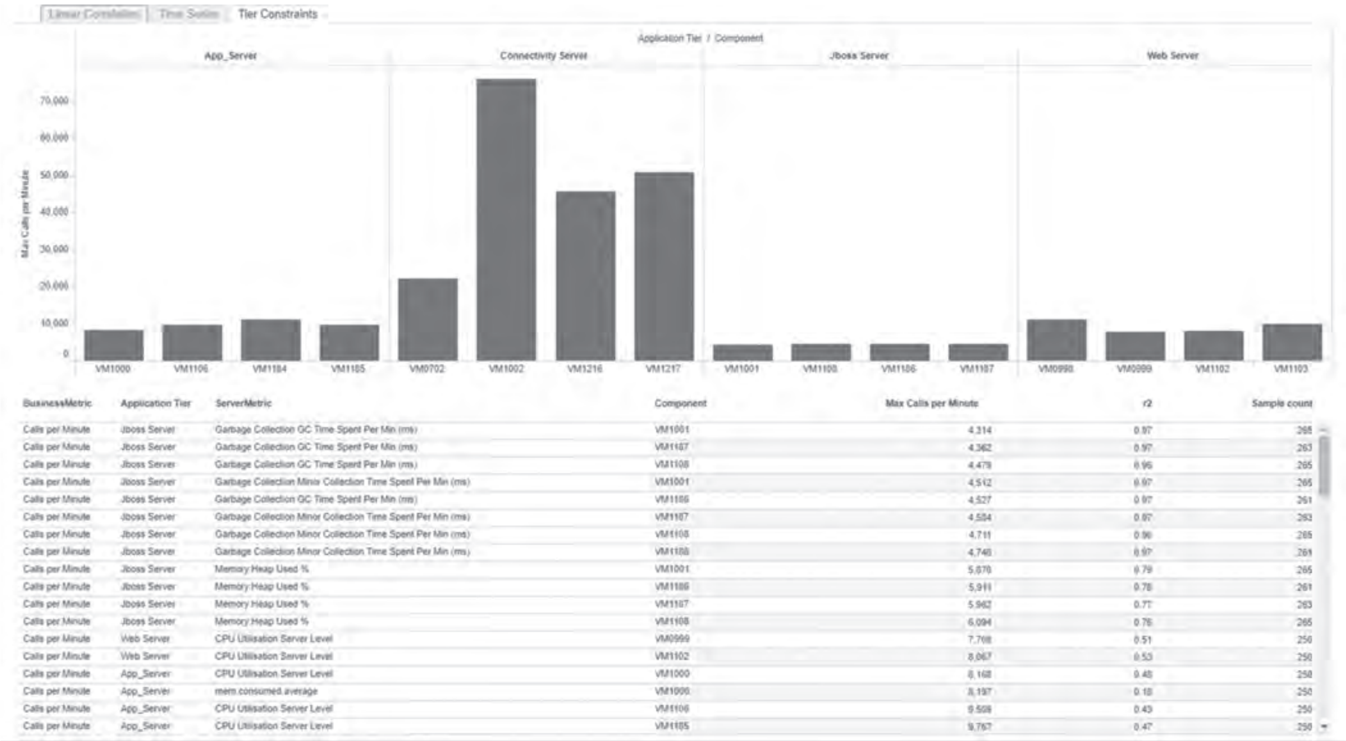
the operations would be successful ones (see Figure 7).

The results were as follows:

- using the application demand modelling process, the bank was able to clear constraints; troubling bottlenecks were also identified;
- no CPU or memory capacity issues resulted from the projected increased load;
- performance problems were detected, which led to an increase in response times;
- the addition of four more VMs to the Jboss layer was expected to resolve the issue; and
- no new infrastructure was required.

For companies looking to implement a capacity management solution, ITRS recommends they do the following:

- identify where the instance needs to run (location) and optimise for cost/performance;



BusinessMetric	Application Tier	ServerMetric	Component	Max Calls per Minute	r2	Sample count
Calls per Minute	Jboss Server	Garbage Collection GC Time Spent Per Min (ms)	VM1001	4,314	0.97	265
Calls per Minute	Jboss Server	Garbage Collection GC Time Spent Per Min (ms)	VM1187	4,362	0.97	263
Calls per Minute	Jboss Server	Garbage Collection GC Time Spent Per Min (ms)	VM1108	4,479	0.96	265
Calls per Minute	Jboss Server	Garbage Collection Minor Collection Time Spent Per Min (ms)	VM1001	4,512	0.97	265
Calls per Minute	Jboss Server	Garbage Collection GC Time Spent Per Min (ms)	VM1186	4,527	0.97	261
Calls per Minute	Jboss Server	Garbage Collection Minor Collection Time Spent Per Min (ms)	VM1187	4,584	0.97	263
Calls per Minute	Jboss Server	Garbage Collection Minor Collection Time Spent Per Min (ms)	VM1108	4,711	0.96	265
Calls per Minute	Jboss Server	Garbage Collection Minor Collection Time Spent Per Min (ms)	VM1188	4,746	0.97	261
Calls per Minute	Jboss Server	Memory Heap Used %	VM1001	5,070	0.79	265
Calls per Minute	Jboss Server	Memory Heap Used %	VM1186	5,911	0.78	261
Calls per Minute	Jboss Server	Memory Heap Used %	VM1187	5,962	0.77	263
Calls per Minute	Jboss Server	Memory Heap Used %	VM1108	6,094	0.76	265
Calls per Minute	Web Server	CPU Utilisation Server Level	VM0999	7,708	0.51	250
Calls per Minute	Web Server	CPU Utilisation Server Level	VM1102	8,067	0.53	250
Calls per Minute	App_Server	CPU Utilisation Server Level	VM1000	8,168	0.48	258
Calls per Minute	App_Server	mem.consumed.average	VM1000	8,197	0.18	250
Calls per Minute	App_Server	CPU Utilisation Server Level	VM1106	9,509	0.43	250
Calls per Minute	App_Server	CPU Utilisation Server Level	VM1185	9,767	0.47	250

Figure 4: Tier constraints
Source: ITRS.

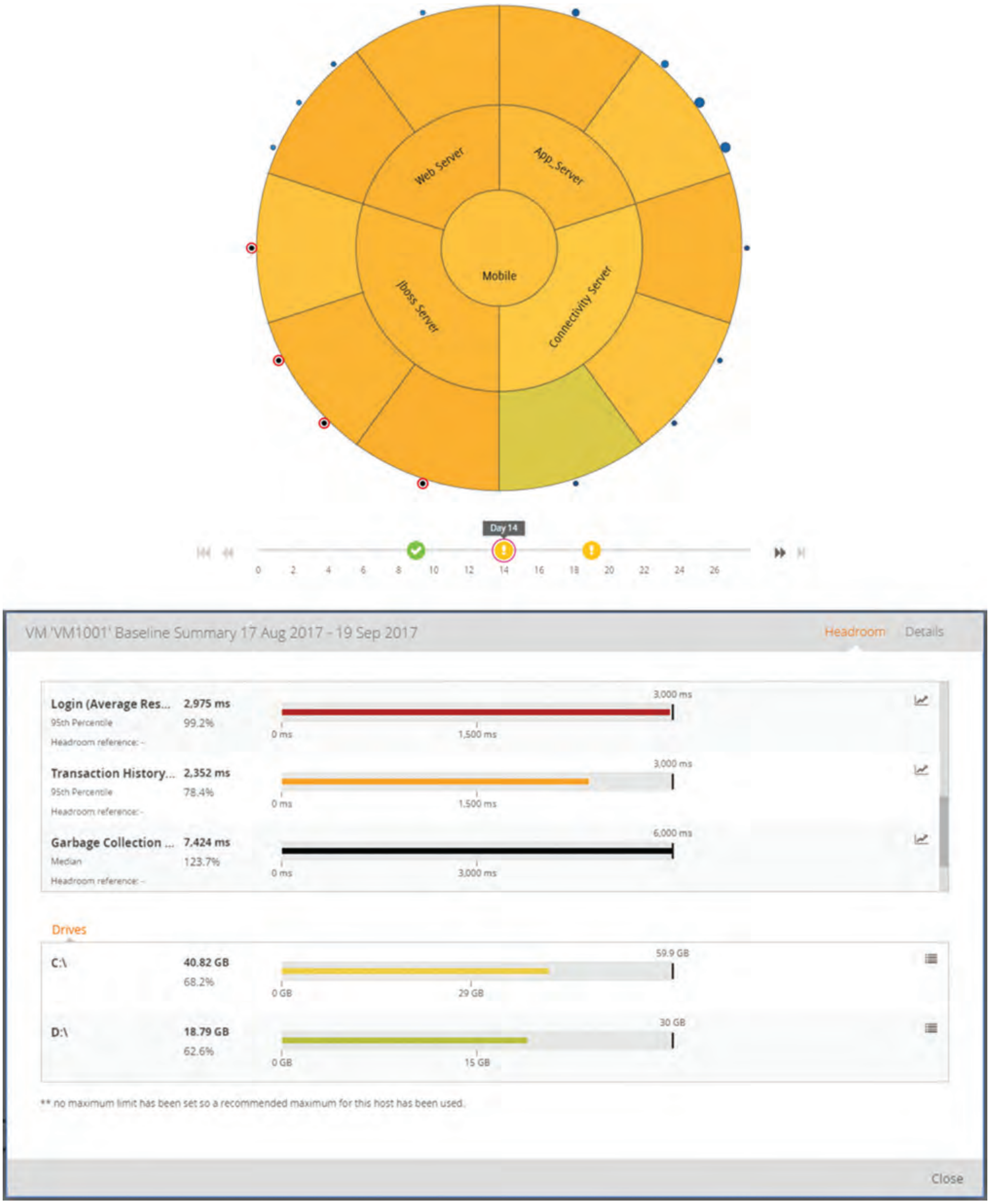
- identify the best way to buy the instance, which depends on how long it is going to run for — this is an aggregated need for that instance size, not the need for one specific instance for one application;
- identify how long an instance should run for, and if it is idle, how long before it should be shut down; and
- analyse the billing engines of the cloud providers on an ongoing basis to identify optimal usage and policies.

Ideally, the capacity planning tool should be an active tool, so that some types of recommendations/policies can be actioned automatically. Finally, the tool should be able to compare multiple cloud providers and then optimise across the hybrid-cloud IT estate.

Additionally, ITRS recommends companies optimise their usage and continuously review the following:

- optimise the cloud at the application level by correlating business demand with cloud service utilisation;
- plan for growth and predict upcoming costs with advanced predictive analytics and forward-thinking what-if scenario modelling;
- improve business processes with service management integration; and
- manage across hybrid-IT, on-prem and multi-cloud in a single tool with consistent reporting regardless of the environment.

Figure 7 shows how right-sizing (ie 'the process of matching instance types and sizes to your workload performance and capacity requirements at the lowest possible cost'[29]) works. As AWS contends, right-sizing is 'also the process of looking at deployed instances and identifying opportunities to eliminate or downsize without compromising capacity or other requirements, which results in



VM 'VM1001' Baseline Summary 17 Aug 2017 - 19 Sep 2017

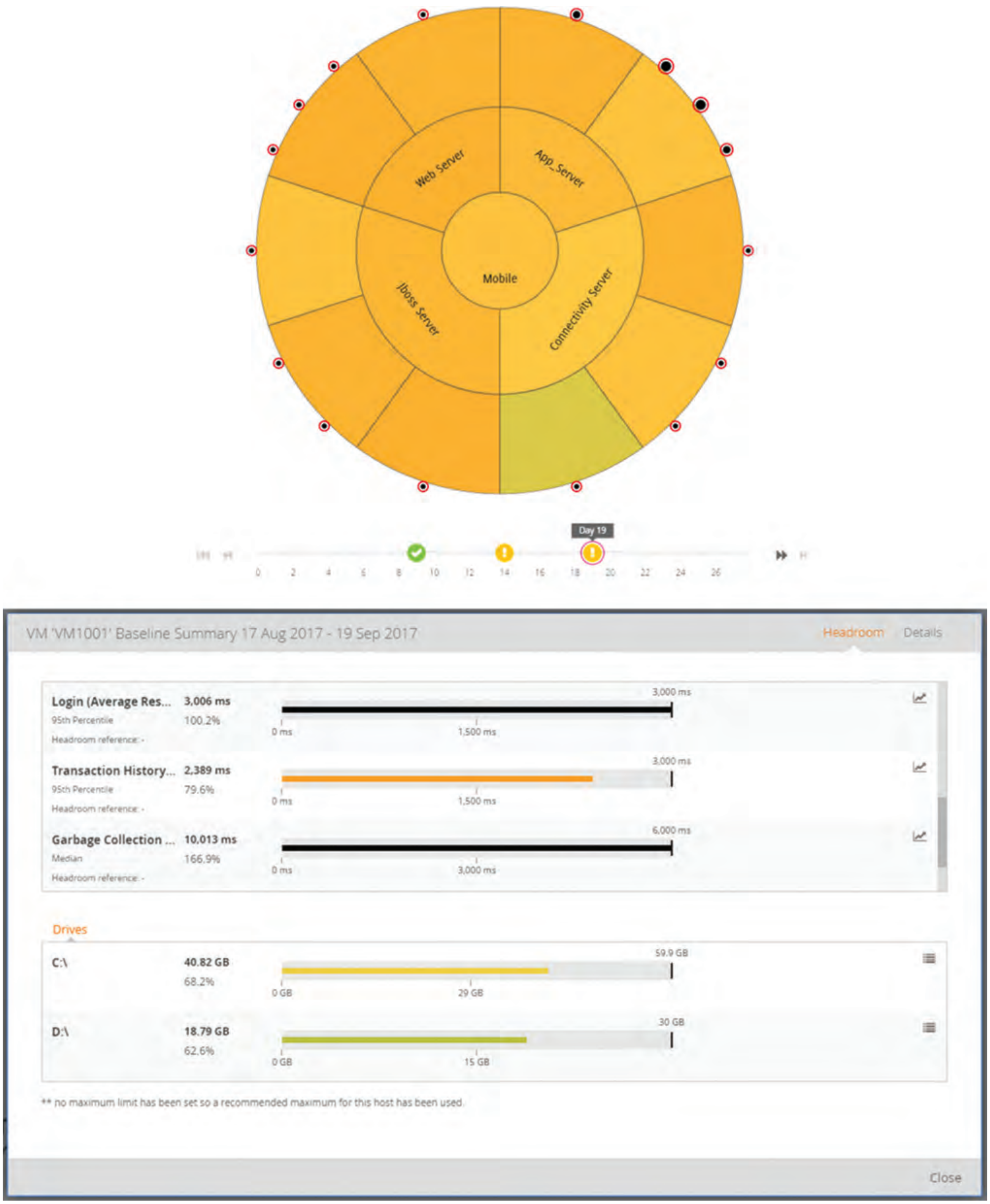
Login (Average Res...	2,975 ms	
95th Percentile	99.2%	
Headroom reference: -		
Transaction History...	2,352 ms	
95th Percentile	78.4%	
Headroom reference: -		
Garbage Collection ...	7,424 ms	
Median	123.7%	
Headroom reference: -		

Drives

C:\	40.82 GB	
	68.2%	
D:\	18.79 GB	
	62.6%	

** no maximum limit has been set so a recommended maximum for this host has been used.

Figure 5: Scenario modelling
Source: ITRS.



VM 'VM1001' Baseline Summary 17 Aug 2017 - 19 Sep 2017

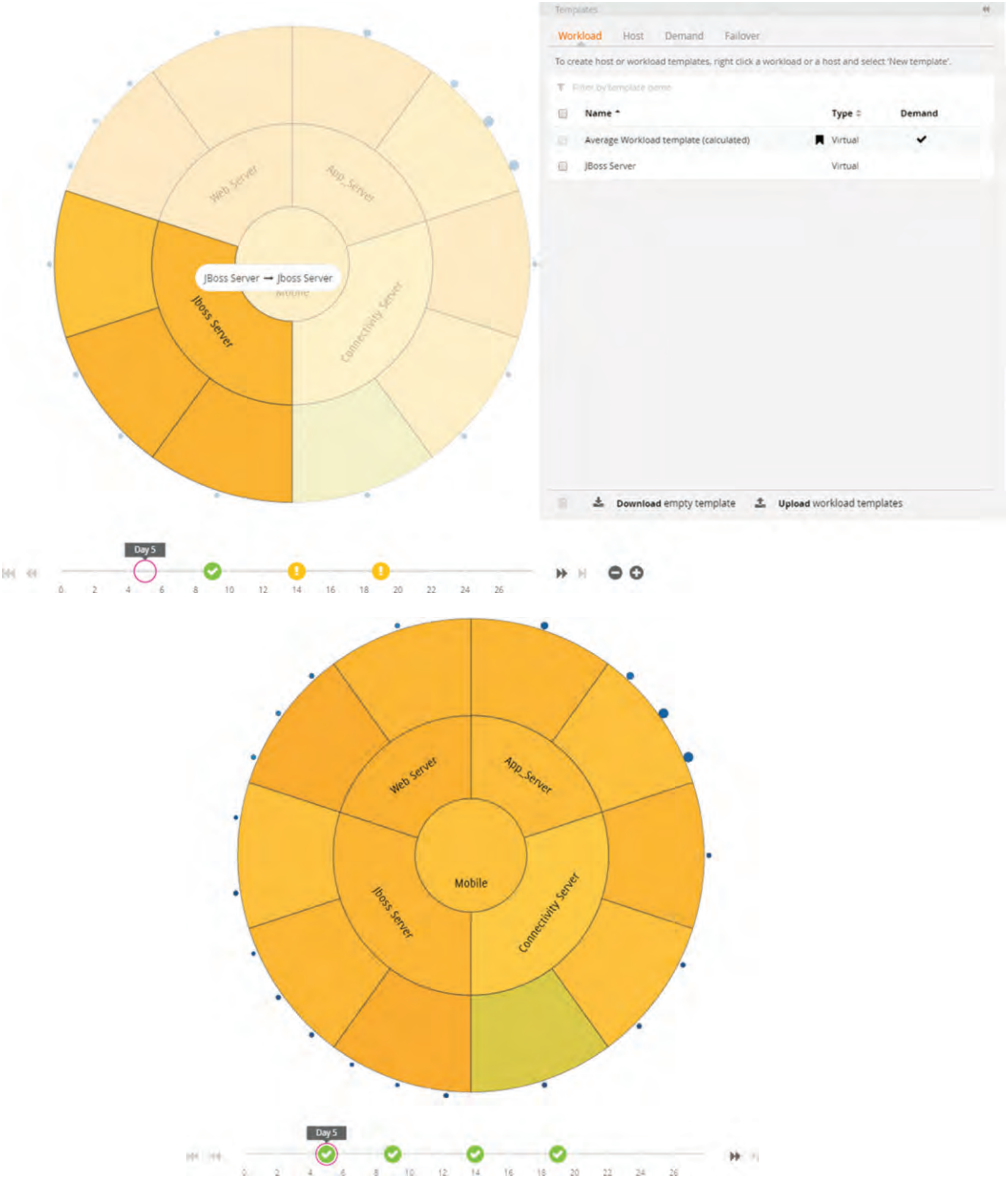
Login (Average Res... 95th Percentile Headroom reference: -	**3,006 ms** 100.2%	0 ms – 1,500 ms – 3,000 ms
Transaction History... 95th Percentile Headroom reference: -	**2,389 ms** 79.6%	0 ms – 1,500 ms – 3,000 ms
Garbage Collection ... Median Headroom reference: -	**10,013 ms** 166.9%	0 ms – 3,000 ms – 6,000 ms

Drives

C:\	**40.82 GB** 68.2%	0 GB – 29 GB – 59.9 GB
D:\	**18.79 GB** 62.6%	0 GB – 15 GB – 30 GB

** no maximum limit has been set so a recommended maximum for this host has been used.

Figure 6: Scenario modelling
Source: ITRS.



Templates

Workload Host Demand Failover

To create host or workload templates, right click a workload or a host and select 'New template'.

Filter by template name

□	Name ˄	Type ⇕	Demand
□	Average Workload template (calculated)	Virtual	✓
□	JBoss Server	Virtual	

Download empty template **Upload** workload templates

Figure 7: Scenario modelling
Source: ITRS.

lower costs'.[29] When moving an application currently running in a data centre to the cloud, one can capture all the information on the servers needed before the actual move to the cloud.

Figure 8 illustrates a data centre running the ITRS Capacity Planner tool. Blue areas are under-utilised while orange areas are more heavily utilised. Capacity Planner shows the recommendations to improve the use of the data centre servers.

However, if the application is to be ported to the cloud as in Figure 9, it makes no sense to buy the extra capacity in the cloud, as it is easy to re-size the system if and when it is needed. So, the tool reveals the cost of moving 'like for like' in case it is important to know what the baseline is. The cost of moving the application right-sized into the cloud is also shown. In this case, a savings of 28 per cent will be attained on a one-year plan.

Most cloud cost optimisation tools work at the total cloud spend level, optimising the entire cloud estate. For some larger companies and their application teams, this is too broad an undertaking to commit to. ITRS Capacity Planner, however, works at the single instance level, so separate modelling and application optimisation can be achieved.

If an on-premises workload is to be moved to the cloud, Capacity Planner prices up the 'like for like' migration and the right-sized estate, identifying cost savings as per Figure 10.

Applications can be analysed and optimised individually or in the aggregate. Certain applications may have different resource usage requirements and therefore not be suitable for aggregation. In this example, high memory and a fast disk are necessities.

The user can define the 'recommendation rules' (see Figure 11) that drive the optimisation and migration of specific application needs. In this case, the required memory and fast disks will be provided. So, when planning the migration of this application, the cloud provisioning team will be informed that 'memory-optimised' instances are required in order to provide 'burstable' dynamic memory and solid-state disks for the fastest disk access possible.

Right-size options are presented visually (see Figure 12), with configurations matched to the closest instance sizes based on statistical profiling of demand. In this example, a client had bought an Oracle Linux 4/5 server, but when the usage of CPU and memory were analysed, an instance that better matched with the statistical profiling of the demand was recommended.

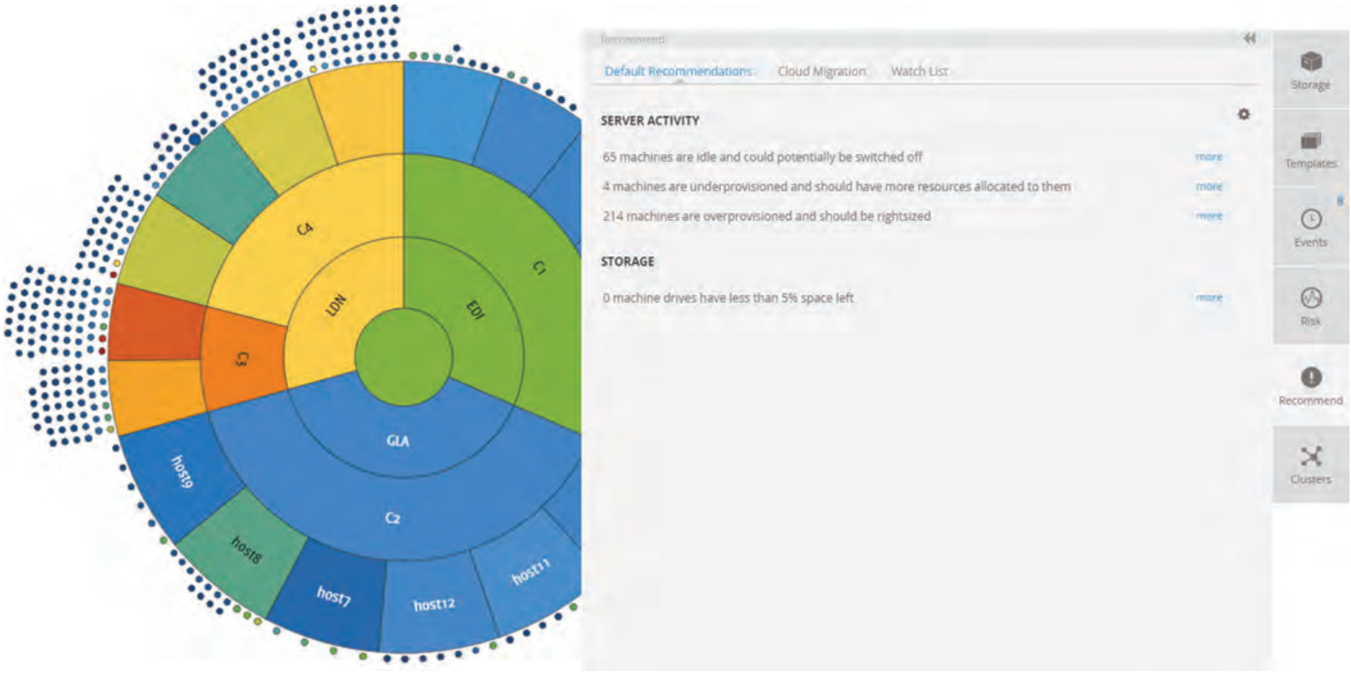
The next stage is right-buying. Once the instances needed to run the application are known, Capacity Planner looks at where and for how long the instances will run.

Figure 13 shows an IT estate that had been left running constantly. The blue areas of the estate are underutilised or idle systems that have not been halted. Running instances when there are no workloads is one of the biggest causes of wastage in cloud spend. Most of the time, many of these unused instances can — and should — be shut down.

Figure 14 shows the same estate but with the Capacity Planning right-buying and policy management in place. The black area reveals instances that have been automatically shut down and then spun back up as and when needed. This activity resulted in a 75 per cent AWS saving. It also makes the entire system much more productive.

ITRS recommends the following actions when implementing the Capacity Planning solution:

- *Right size:*
  - obtain a highly granular data capture of all resource usage (CPU, memory, disk, network) and identify the sizes of instances needed for each application workload;



SERVER ACTIVITY

65 machines are idle and could potentially be switched off — more

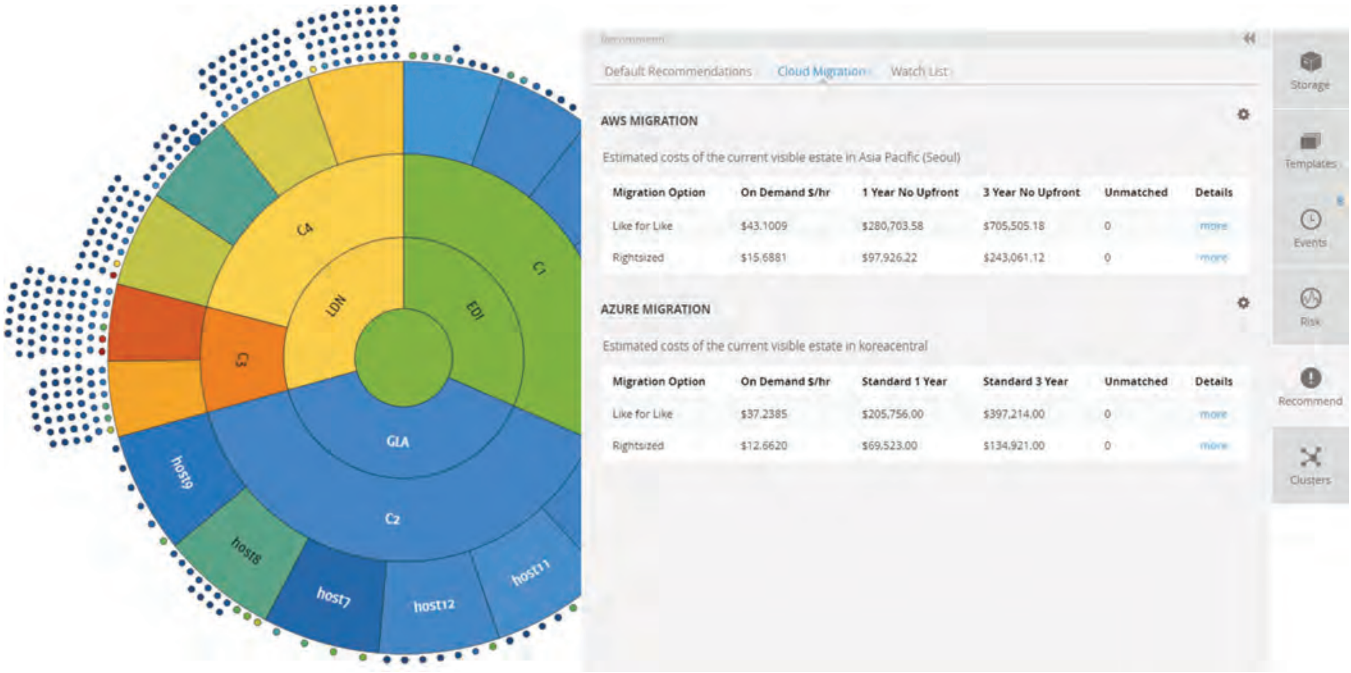
4 machines are underprovisioned and should have more resources allocated to them — more

214 machines are overprovisioned and should be rightsized — more

STORAGE

0 machine drives have less than 5% space left — more

Figure 8: On-premise recommendations
Source: ITRS.



AWS MIGRATION

Estimated costs of the current visible estate in Asia Pacific (Seoul)

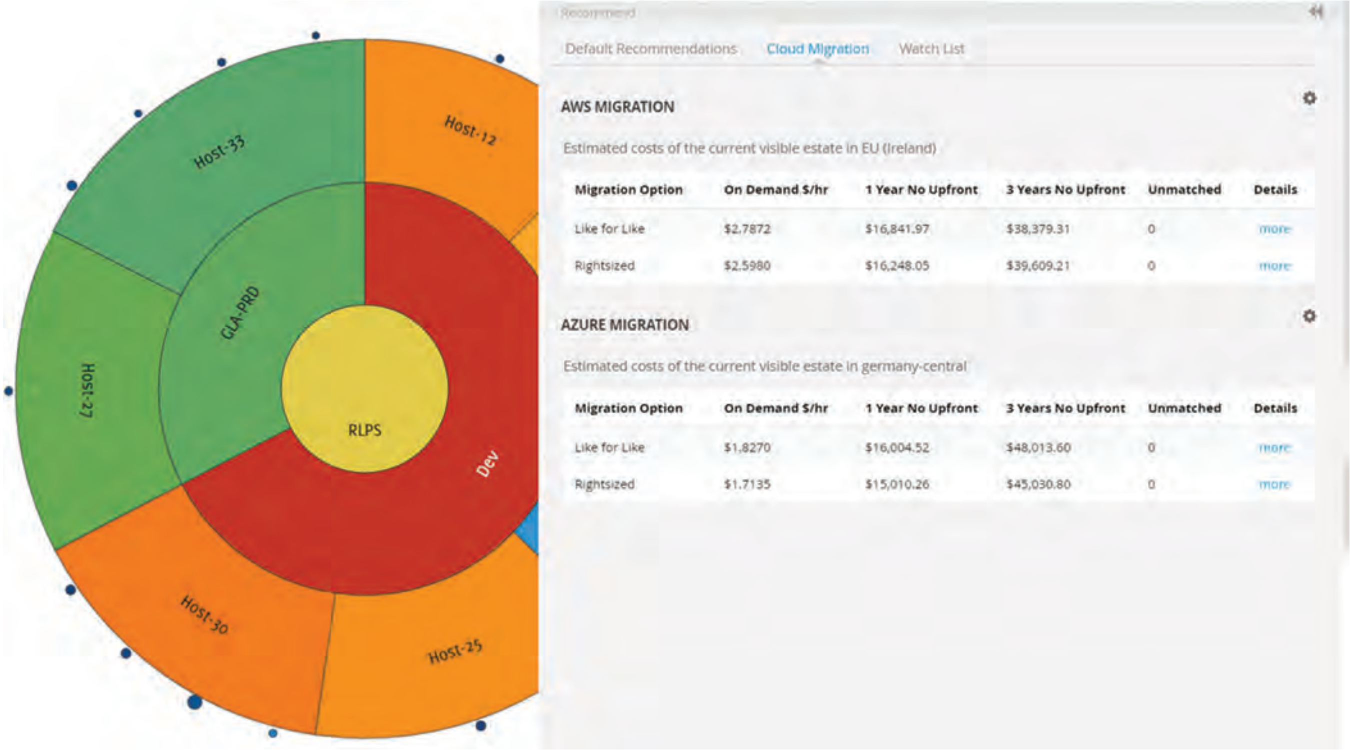
Migration Option	On Demand $/hr	1 Year No Upfront	3 Year No Upfront	Unmatched	Details
Like for Like	$43.1009	$280,703.58	$705,505.18	0	more
Rightsized	$15.6881	$97,926.22	$243,061.12	0	more

AZURE MIGRATION

Estimated costs of the current visible estate in koreacentral

Migration Option	On Demand $/hr	Standard 1 Year	Standard 3 Year	Unmatched	Details
Like for Like	$37.2385	$205,756.00	$397,214.00	0	more
Rightsized	$12.6620	$69,523.00	$134,921.00	0	more

Figure 9: Migration from on-premises workload to the cloud
Source: ITRS.



AWS MIGRATION

Estimated costs of the current visible estate in EU (Ireland)

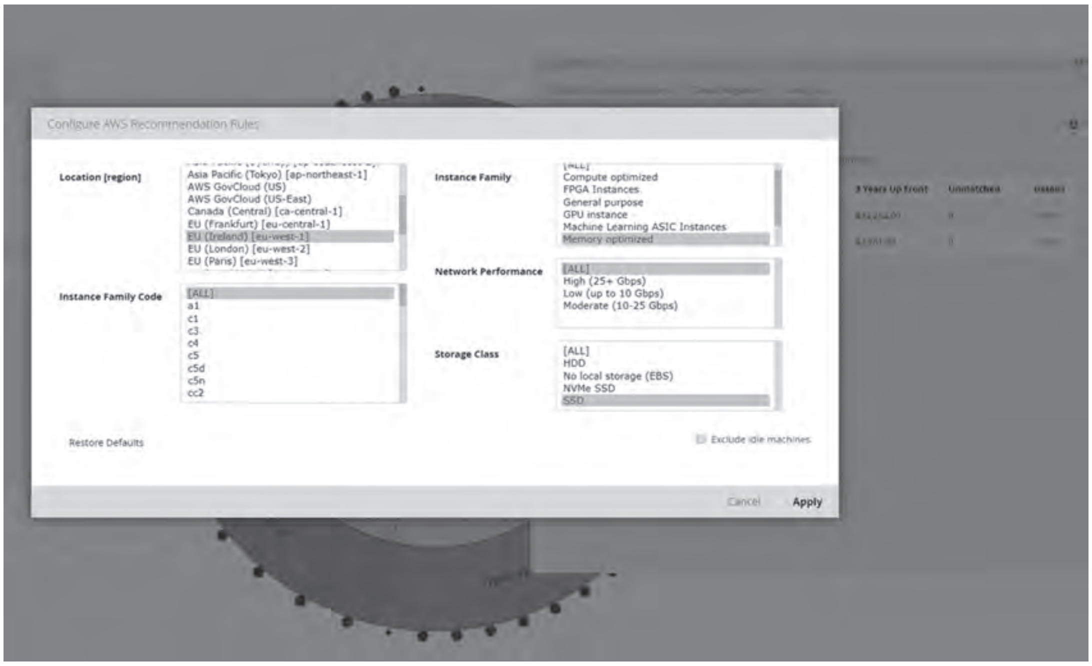
Migration Option	On Demand $/hr	1 Year No Upfront	3 Years No Upfront	Unmatched	Details
Like for Like	$2.7872	$16,841.97	$38,379.31	0	more
Rightsized	$2.5980	$16,248.05	$39,609.21	0	more

AZURE MIGRATION

Estimated costs of the current visible estate in germany-central

Migration Option	On Demand $/hr	1 Year No Upfront	3 Years No Upfront	Unmatched	Details
Like for Like	$1.8270	$16,004.52	$48,013.60	0	more
Rightsized	$1.7135	$15,010.26	$45,030.80	0	more

Figure 10: Adding application-level visibility
Source: ITRS.



Configure AWS Recommendation Rules

Location [region]
Asia Pacific (Tokyo) [ap-northeast-1]
AWS GovCloud (US)
AWS GovCloud (US-East)
Canada (Central) [ca-central-1]
EU (Frankfurt) [eu-central-1]
EU (Ireland) [eu-west-1]
EU (London) [eu-west-2]
EU (Paris) [eu-west-3]

Instance Family Code
[ALL]
a1
c1
c3
c4
c5
c5d
c5n
cc2

Instance Family
[ALL]
Compute optimized
FPGA Instances
General purpose
GPU instance
Machine Learning ASIC Instances
Memory optimized

Network Performance
[ALL]
High (25+ Gbps)
Low (up to 10 Gbps)
Moderate (10-25 Gbps)

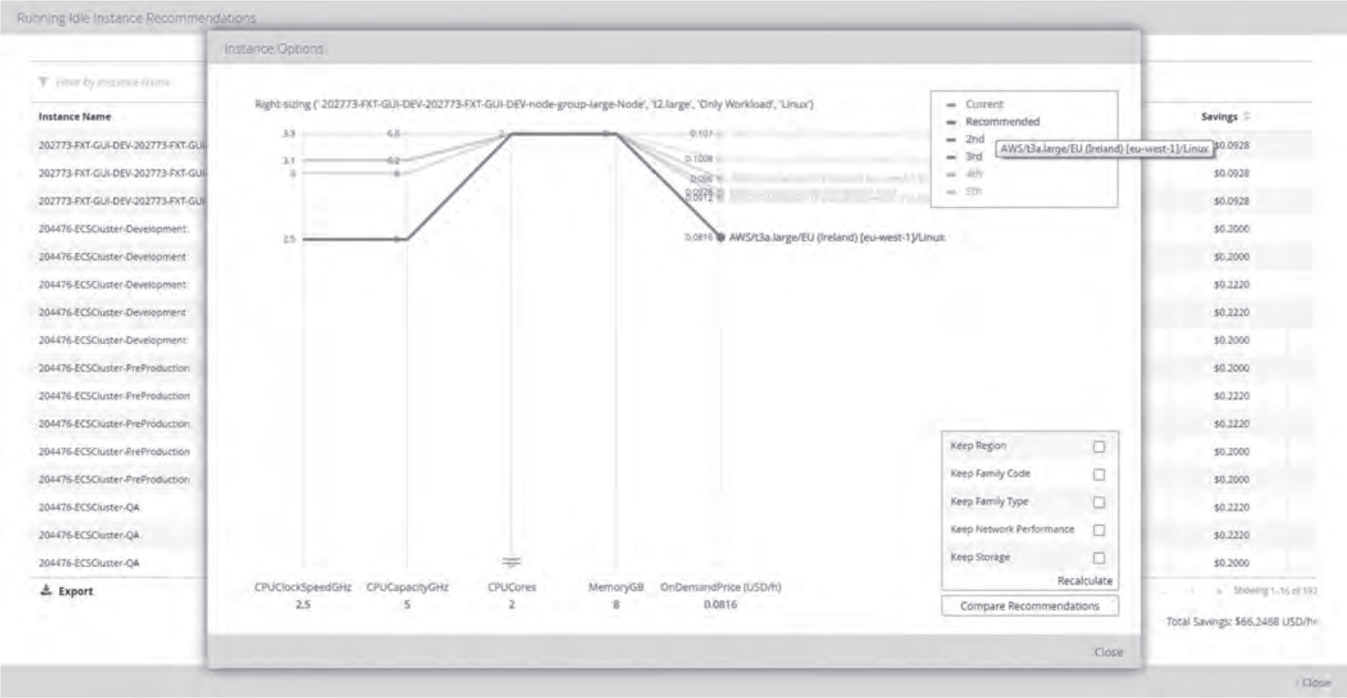
Storage Class
[ALL]
HDD
No local storage (EBS)
NVMe SSD
SSD

Restore Defaults

□ Exclude idle machines

Cancel **Apply**

Figure 11: Configuring recommendation rules
Source: ITRS.



Running Idle Instance Recommendations

Instance Options

Right-sizing (' 202773-FXT-GUI-DEV-202773-FXT-GUI-DEV-node-group-large-Node', 't2.large', 'Only Workload', 'Linux')

Instance Name	Savings
202773-FXT-GUI-DEV-202773-FXT-GUI	$0.0928
202773-FXT-GUI-DEV-202773-FXT-GUI	$0.0928
202773-FXT-GUI-DEV-202773-FXT-GUI	$0.0928
204476-ECSCluster-Development	$0.2000
204476-ECSCluster-Development	$0.2000
204476-ECSCluster-Development	$0.2220
204476-ECSCluster-Development	$0.2220
204476-ECSCluster-Development	$0.2000
204476-ECSCluster-PreProduction	$0.2000
204476-ECSCluster-PreProduction	$0.2220
204476-ECSCluster-PreProduction	$0.2220
204476-ECSCluster-PreProduction	$0.2000
204476-ECSCluster-PreProduction	$0.2000
204476-ECSCluster-QA	$0.2220
204476-ECSCluster-QA	$0.2220
204476-ECSCluster-QA	$0.2000

CPUClockSpeedGHz	CPUCapacityGHz	CPUCores	MemoryGB	OnDemandPrice (USD/h)
2.5	5	2	8	0.0816

Keep Region □
Keep Family Code □
Keep Family Type □
Keep Network Performance □
Keep Storage □

Showing 1–16 of 193

Total Savings: $66.2468 USD/h

Figure 12: Configuring recommendation rules
Source: ITRS.

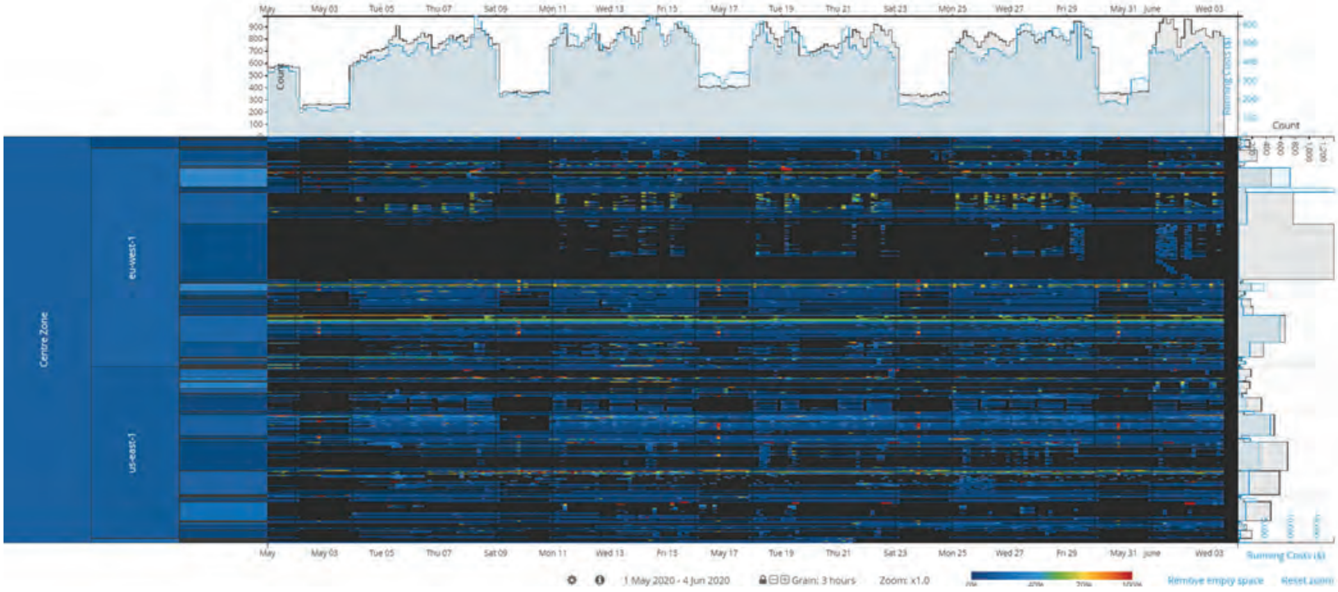


Figure 13: Timeburst view
Source: ITRS.

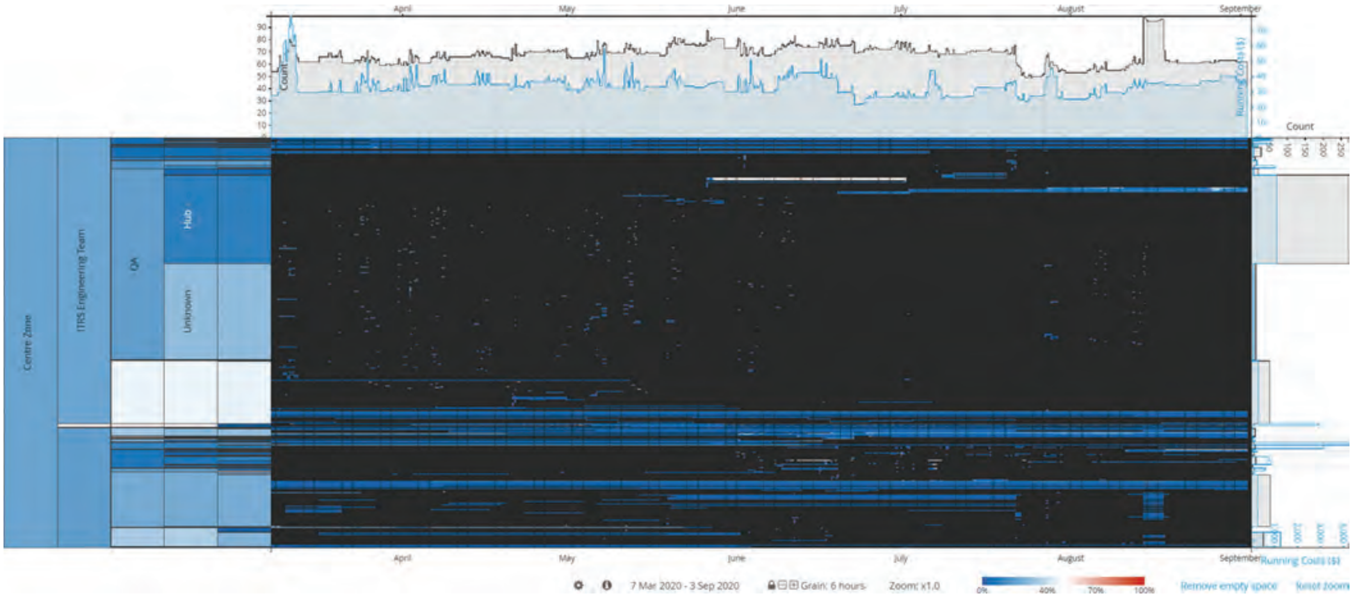


Figure 14: Timeburst view showing that most workloads are very short-lived with minimal idle time
Source: ITRS.

  - determine optimum configuration of burstable or non-burstable instances;
  - identify idle times and workload periodicity; and
  - understand the capacity of hybrid-IT and repatriate work from the cloud if on-premises capacity allows it.
- *Right buy:*
  - identify where the instance needs to run (location) and optimise for cost/performance;
  - identify the best way to buy the instance based on how long it will run;
  - identify how long an instance should run for, and if it is idle, how long before it should be shut down; and
  - analyse the billing engines of the cloud providers on an ongoing basis to identify optimal usage and policies.
- *Optimise:*
  - right-size at the application level by correlating business demand with cloud service utilisation;
  - plan for growth and predict upcoming costs with advanced predictive analytics and forward-thinking 'what-if' scenario modelling;
  - improve business processes with service management integration; and
  - manage across hybrid-IT, on-prem and multi-cloud in a single tool with consistent reporting regardless of the environment.

CONCLUSION

As Lee argues, 'capacity planning is a challenging task when there is an unpredictable, fluctuating computing demand with many peaks and troughs'.[13] Fortunately, not only are there many ways to track capacity data, but analytics can produce incredibly accurate evaluation models to help marketers get fast returns from their capacity planning and marketing initiatives.

Capacity planning works hand-in-hand with Kalb's condition that every dollar spent on marketing be valued and quantified.[3] It seeks to minimise wastage when it comes to the IT estate, including marketing's IT activities. The question is, can marketing's IT activities be quantified so that every sale, lead and marketing offer can be traced back to the marketing effort that produced it? This a lofty goal but IT is getting there.

ITRS recommends companies optimise and continuously review their cloud usage at the application level by correlating business demand with cloud service utilisation. Amazon became one of the most valuable companies in the world on the back of the profits it made on its AWS service rather than its marketplace, where margins are slim. This should be a warning to anyone planning to move their IT services into the cloud. Although it might make a lot of budgetary sense to go to the cloud, there is no reason to pay for servers that shouldn't be on.

Companies should plan for growth and predict upcoming costs with advanced predictive analytics and forward-thinking what-if scenario modelling. COVID-19 came out of nowhere and threw a wrench into a lot of business forecasting models, but that was a unique situation and it spurred a work-for-home trend that put great strains on IT departments the world over. However, it also showed the value of the cloud and ITRS's recommendation that companies should manage their hybrid-IT, on-prem and multi-cloud platforms through a single tool to provide consistent reporting regardless of the environment. This should help businesses contain costs.

While some things have improved since John Naisbitt's warning that 'we are drowning in information but starved for knowledge',[4] we are still overwhelmed by data. That said, one can now capture, track and utilise data so much more easily than was possible only a few short years ago. Numerous products in the market today support the kind of tracking that provides the baseline for change. With the IoT revolution about to hit with massive new amounts of harvested data coming online, businesses must understand that the cloud can be a good place to find help, but without adequate control, costs can soon spiral out of control.

Rather than waiting for a dream solution to materialise, marketers must find a system that enables them to make better decisions and they need to support those decisions with verifiable data. Wanamaker's famous lament that he did not know which half of his marketing spend was wasted no longer rings true. Today, the dominant advertising, social media, customer relationship management and marketing software companies are focused on marketing attribution, trying to help companies quantify every dollar they spend on advertising. This is an important part of the equation, but it is just as important to understand the costs to attain that attribution. In the words of the consumer behaviour specialist Steuart Henderson Britt, 'doing business without advertising is like winking at a girl in the dark. You know what you are doing, but nobody else does'.[30] This pretty much exemplifies what goes on behind the scenes with capacity planning, real-time monitoring, AIOps, cloud bursting and predictive resource scaling. It is a lot, it is incredibly valuable, and it is something marketing IT departments should team up to address. Few people realise that a strong operations backbone can be the impetus that helps marketing create the wink that might just entice the client enough to close the sale.

References

1. DuBois, J. (n.d.) 'What is a constraint in marketing?', available at: https://smallbusiness.chron.com/constraint-marketing-65978.html (accessed 14th December, 2020).
2. Goldratt, E.M. (2004) 'The Goal: a Process of Ongoing Improvement', North River Press, Great Barrington, MA.
3. Kalb, I. (2013) '8 steps to creating an effective marketing information system', *Business Insider*, 22nd November, available at: https://www.businessinsider.com/the-marketing-information-system-the-missing-link-for-greater-success-2013-11 (accessed 13th December, 2020).
4. Naisbitt, J. (1982) 'Megatrends: Ten New Directions Transforming Our Lives', Warner Books, Inc, New York, NY.
5. Flexera (2020) 'Flexera 2020 State of the Cloud Report', available at: https://info.flexera.com/SLO-CM-REPORT-State-of-the-Cloud-2020 (accessed 6th November, 2020).

6. Kotler, P. (1988) ‘Marketing Management: Analysis’, Planning and Control’, Prentice-Hall, Upper Saddle River, NJ.
7. Sebestyén, Z. and Juhász, V. (2003) ‘The impact of the cost of unused capacity on production planning of flexible manufacturing systems’, *Periodica Polytechnica — Social and Management Sciences*, Vol. 11, No. 2, pp. 185–200.
8. Brynjolfsson, E. (2003) ‘ROI valuation, the IT productivity gap’, available at: https://www.academia.edu/2662751/ROI_Valuation_The_IT_Productivity_GAP (accessed 5th November, 2020).
9. Plant and Works Engineering (2015) ‘10 rules for condition monitoring’, available at: https://pwemag.co.uk/news/fullstory.php/aid/1764/10_rules_for_condition_monitoring.html (accessed 5th November, 2020).
10. Lerner, A. (2017) ‘AIOps platforms’, 9th August, available at: https://blogs.gartner.com/andrew-lerner/2017/08/09/aiops-platforms/ (accessed 14th December, 2020).
11. Fitz, T. (2019) ‘The state of AIOps: Understanding the difference between tools’, available at: https://vmblog.com/archive/2019/10/01/the-state-of-aiops-understanding-the-difference-between-tools.aspx (accessed 14th December, 2020).
12. Soneja, A. (2020) ‘How AIOps is already transforming IT’, CIO.com, 9th March, available at: https://cio.economictimes.indiatimes.com/news/next-gen-technologies/how-aiops-is-already-transforming-it/74544705 (accessed 14th December, 2020).
13. Lee, I. (2017) ‘Determining an optimal mix of hybrid cloud computing for enterprises’, in: ‘Proceedings of the 10th International Conference on Utility and Cloud Computing, Austin, TX, 5th–8th December’, pp. 53–58.
14. Araujo, J., Maciel, P., Andrade, E., Callou, G., Alves, V. and Cunha, P. (2018) ‘Decision making in cloud environments: an approach based on multiple-criteria decision analysis and stochastic models’, *Journal of Cloud Computing, Advanced Systems and Applications*, Vol. 7, No. 7, available at: https://doi.org/10.1186/s13677-018-0106-7. (accessed 26th January, 2021).
15. López-Pires, F. and Barán, B. (2017) ‘Cloud computing resource allocation taxonomies’, *International Journal of Cloud Computing*, Vol. 6, No. 3, pp. 238–264.
16. Balaji, M., Kumar, A. and Rao, S.V.R.K. (2018) ‘Predictive cloud resource management framework for enterprise workloads,’ *Journal of King Saud University – Computer and Information Sciences*, Vol. 30, No. 3, pp. 404–415.
17. Wang, C.F., Hung, W.Y. and Yang, C.S. (2014) ‘A prediction based energy conserving resources allocation scheme for cloud computing’, in: ‘Proceedings of the 2014 IEEE International Conference on Granular Computing, Noboribetsu, 24th October’, pp. 320–324.
18. Han, Y., Chan, J. and Leckie, C. (2013) ‘Analysing virtual machine usage in cloud computing’, in: ‘Proceedings of the 2013 IEEE International Conference on Services Computing, Santa Clara, CA, 28th June–3rd July’, pp. 370–377.
19. Deng, D., Lu, Z., Fang, W. and Wu, J. (2013) ‘CloudStreamMedia: a cloud assistant global video on demand leasing scheme’, in: ‘Proceedings of the 2013 IEEE International Conference on Services Computing, Santa Clara, CA, 28th June–3rd July’, pp. 486–493.
20. Laatikainen, G., Mazhelis, O. and Tyrvainen, P. (2016) ‘Cost benefits of flexible hybrid cloud storage: mitigating volume variation with shorter acquisition cycle’, *Journal of System Software*, Vol. 122, pp. 180–201.
21. Guo, T., Sharma, U., Shenoy, P., Wood, T. and Sahu, S. (2014) ‘Cost-aware cloud bursting for enterprise applications’, *ACM Transactions on Internet Technology*, Vol. 13, No. 3, pp. 10:1–10:24.
22. Clemente-Castelló, F.J., Mayo, R. and Fernández, J.C. (2017) ‘Cost model and analysis of iterative MapReduce applications for hybrid cloud bursting’, in: ‘Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Madrid, 14th–17th May’, pp. 858–864.
23. Weinman, J. (2016) ‘Hybrid cloud economics’, *IEEE Cloud Computing*, Vol. 3, No. 1, pp. 18–22.
24. Toosi, A.N., Sinnott, R.O. and Buyya, R. (2018) ‘Resource provisioning for data-intensive applications with deadline constraints on hybrid clouds using Aneka’, *Future Generation Computer Systems*, Vol. 79, Part 2, pp. 765–775.
25. Edmonds, A., Metsch, T., Papaspyrou, A. and Richardson, A. (2012) ‘Toward an open cloud standard’, *IEEE Internet Computing*, Vol. 16, No. 4, pp. 15–25.
26. Forrester (2018) ‘Predictions 2019: Cloud computing comes of age as the foundation for enterprise digital transformation’, available at: https://go.forrester.com/blogs/predictions-2019-cloud-computing/ (accessed 26th January, 2021).
27. Stefanova, H. (2014) ‘Uncovering the hidden lessons in “The Goal” (part 1)’, available at: https://blogs.3ds.com/delmia/uncovering-hidden-lessons-goal-part-1/ (accessed 30th December, 2020).
28. VMWare (2018) ‘Performance Best Practices for VMware vSphere 6.7’, available at: https://docs.vmware.com/en/vRealize-Operations-Manager/6.6/com.vmware.vcom.core.doc/GUID-AEB32BB2-7828-4664-A81A-5E7E3CF38620.html (accessed 5th November, 2020).
29. Amazon (n.d.) ‘Right Sizing’, available at: https://aws.amazon.com/aws-cost-management/aws-cost-optimization/right-sizing/ (accessed 5th November, 2020).
30. Britt, S.H. (1956) *New York Herald Tribune*, 30th October, available at: https://www.oxfordreference.com/view/10.1093/acref/9780191826719.001.0001/q-oro-ed4-00017262 (accessed 26th January, 2020)