# Main challenges :
# Scalability , Availability and Resource Management

Asst. Prof. Ashwini Kumar Mathur

# Scalability Challenge

Hidden Challenges to “**growing**” a system

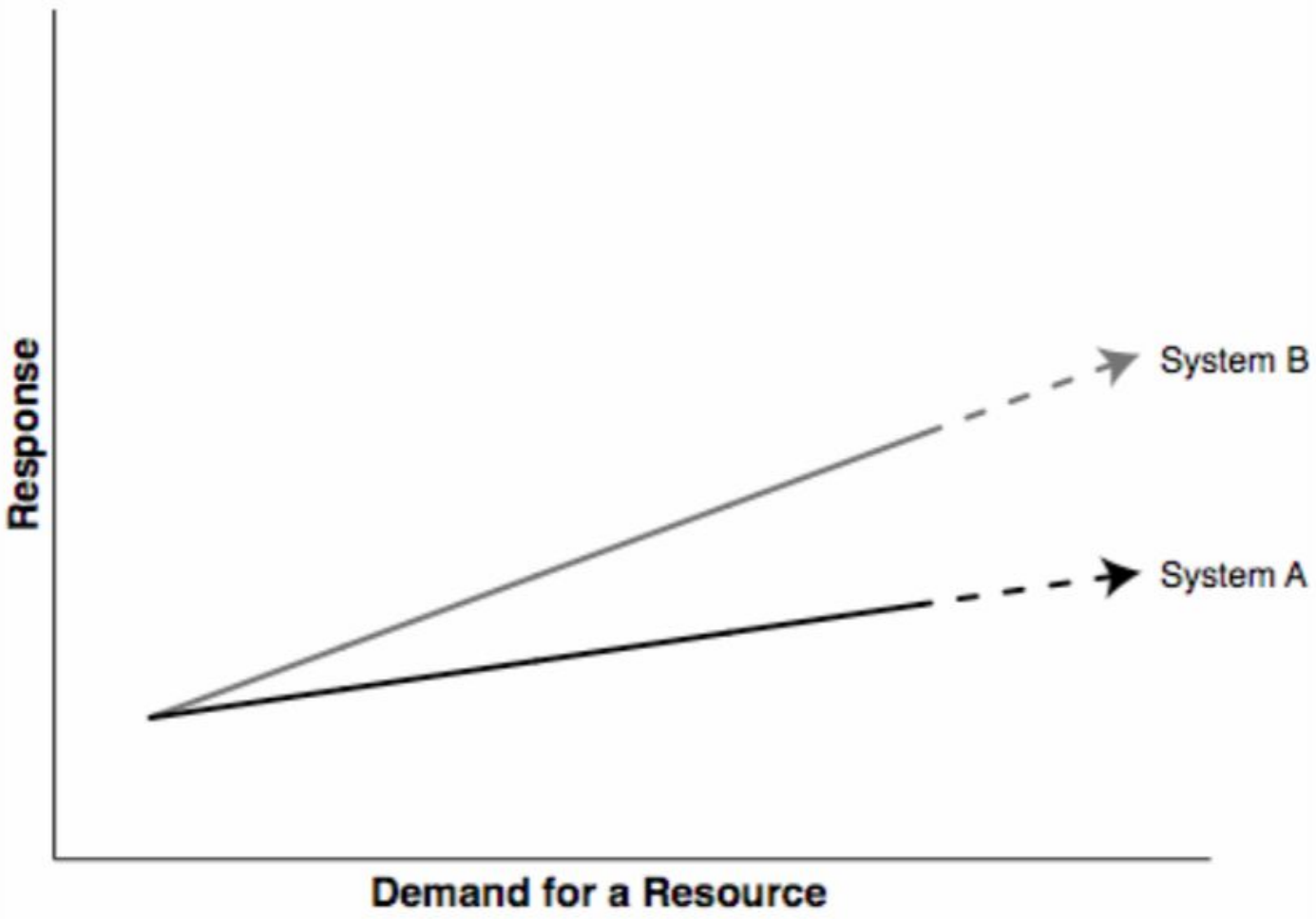
# Let’s Understand first the term “Scalability”

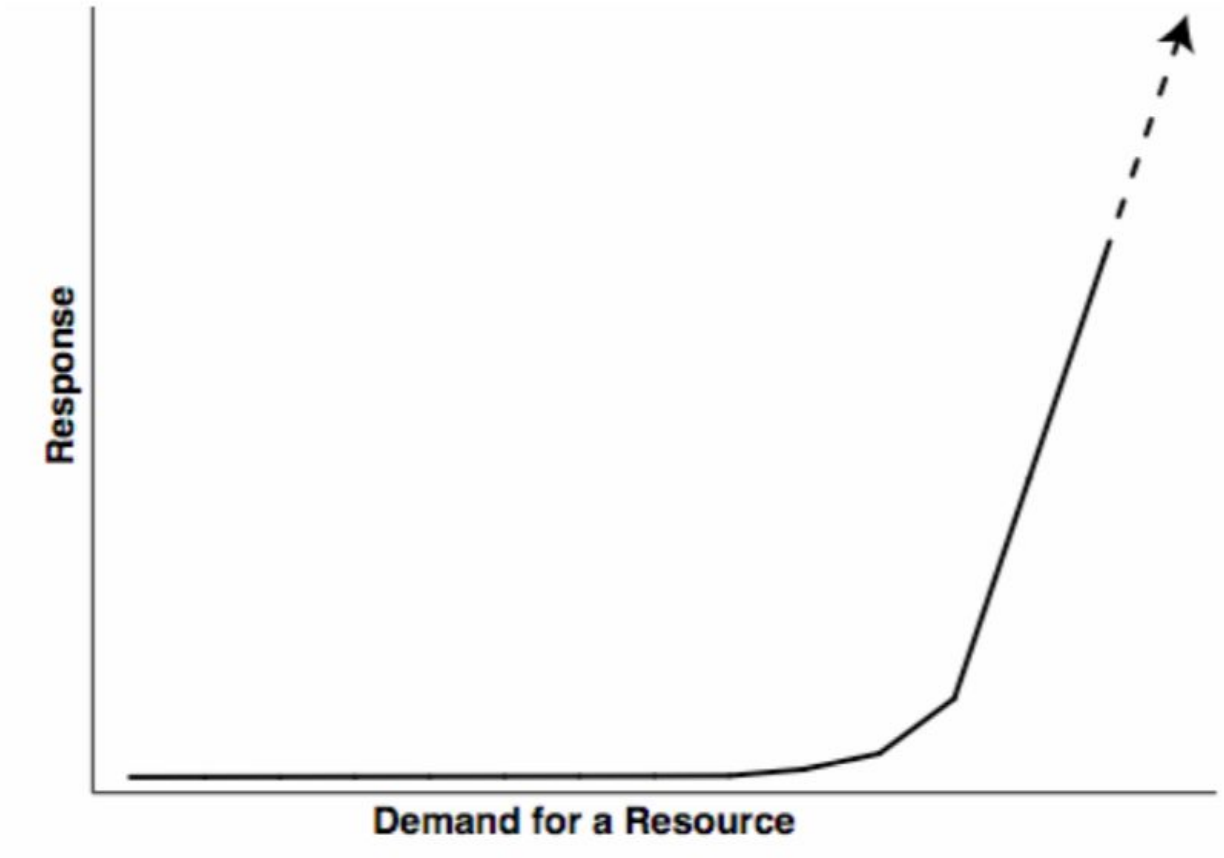
Note: Scalability refers here to the data administration difficulties in creating and maintaining large systems

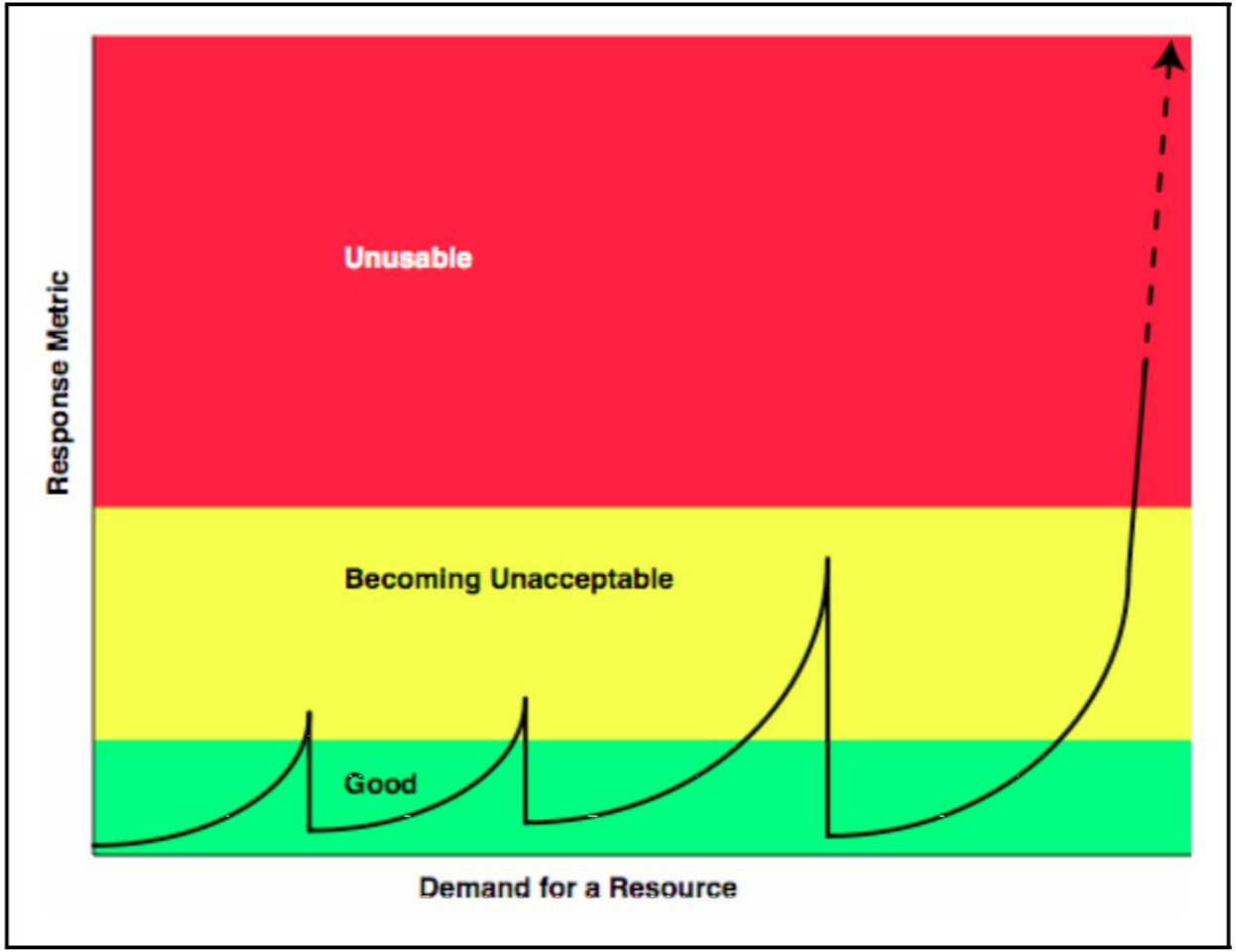
# Overview

Scalability is the **ability to handle increased workload** ( without adding resources to a system ).

Or Scalability is the ability to handle increased workload by repeatedly **applying a cost effective strategy for extending a system’s capacity**.


Response
System B
System A
Demand for a Resource


Response
Demand for a Resource


Unusable
Becoming Unacceptable
Good
Response Metric
Demand for a Resource

# Causes of Scalability Failure

Scalability$_1$ failures occur when increased demand causes some resource to become overloaded or exhausted. This result can be seen in examples where

- Available address space is exceeded.
- Memory is overloaded.
- Available network bandwidth is exceeded.
- An internal table is filled.

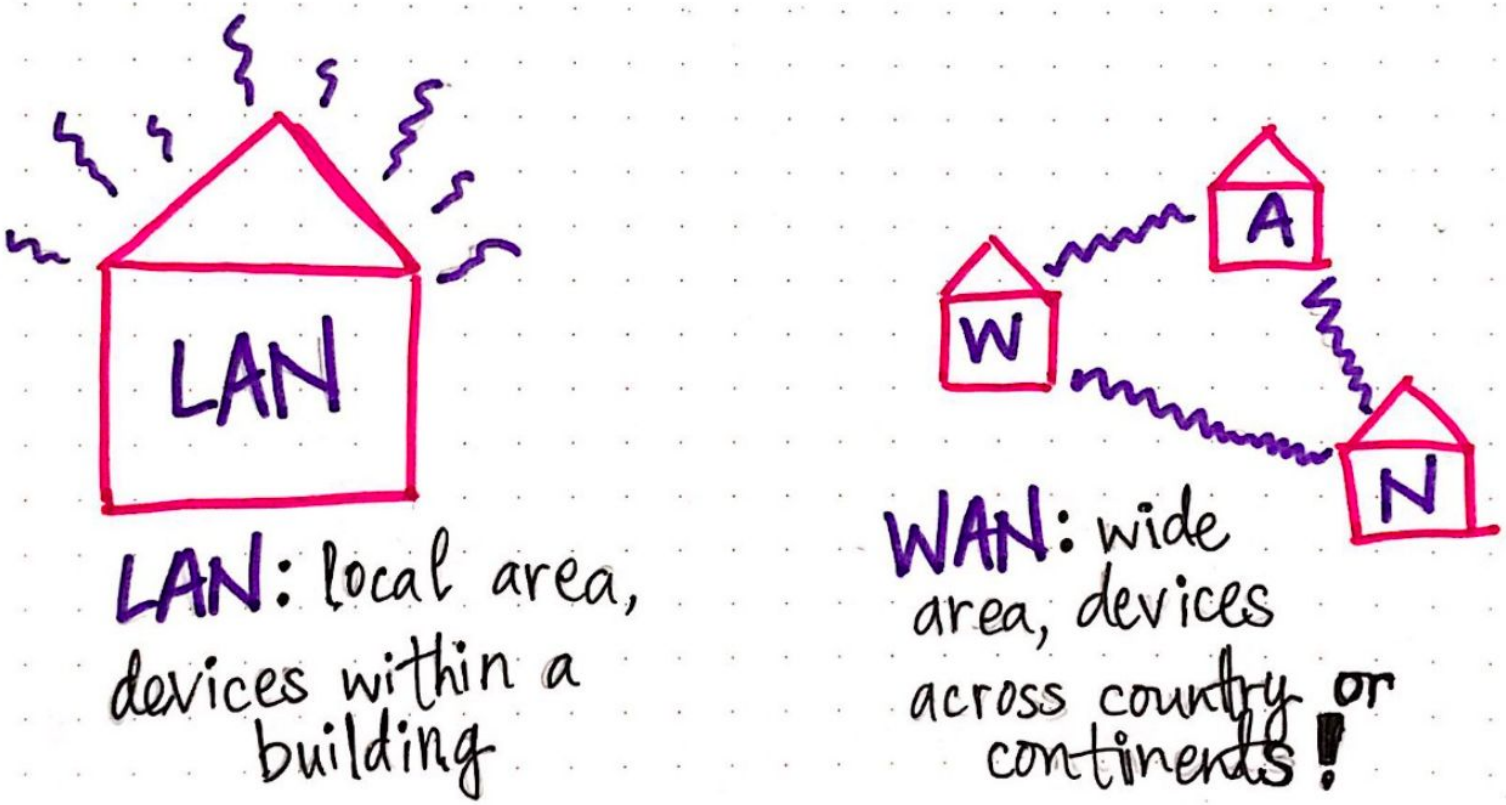
Scalability$_2$ failures occur when some resource is overloaded or exhausted and adding capacity to the resource does not result in a commensurate ability to handle significant additional demand. For example, adding a processor may not allow a system to meet the additional demand if adding the processor also increases overhead significantly.

# #3 Dimension of Scalability

Three **Dimensions** of scalability:

*size* scalability,

*geographical* scalability, and *administrative* scalability




Local vs. wide area networks.

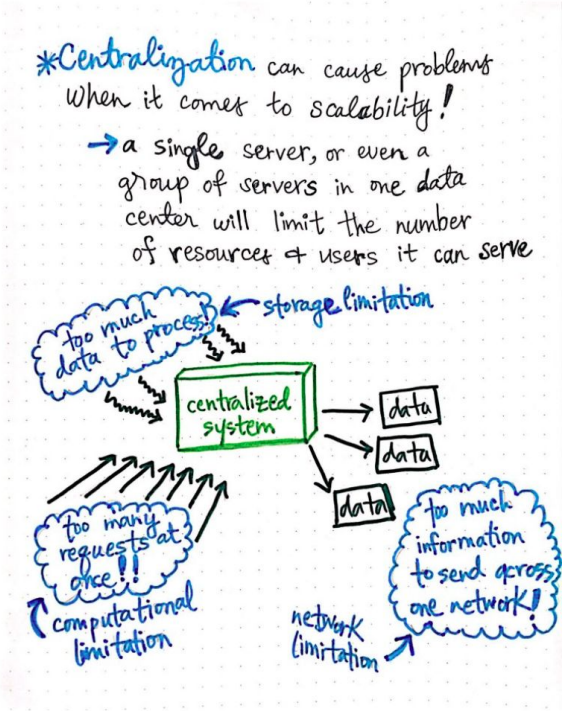
Source : https://medium.com/baseds/scalability-problems-hidden-challenges-of-growing-a-system-f74313b063c3


faced when scaling:

① **Centralization**: everything "lives" in / is controlled by one place.

② **Synchronous communication**: when one service waits around ("blocks") until it receives a reply from another service.

Source : https://medium.com/baseds/scalability-problems-hidden-challenges-of-growing-a-system-f74313b063c3




When centralization makes scaling hard!

* Growing a network that uses synchronous communication can cause problems with scalability!

→ when all servers are on a local network, waiting for one process to complete is fast; but on a wider network this is slower!


When synchronous communication on a (wide-area) network makes scaling hard!

Next challenge is availability

# High Availability

https://www.vmware.com/in/products/vsphere/high-availability.html

# High Availability

## VMware vSphere High Availability

VMware vSphere High Availability delivers the availability required by most applications running in virtual machines, independent of the operating system and applications running in it. High Availability provides uniform, cost-effective failover protection against hardware and operating system outages within your virtualized IT environment. High Availability allows you to:

- Monitor VMware vSphere hosts and virtual machines to detect hardware and guest operating system failures.
- Restart virtual machines on other vSphere hosts in the cluster without manual intervention when a server outage is detected.
- Reduce application downtime by automatically restarting virtual machines upon detection of an operating system failure.

Waiting Room :

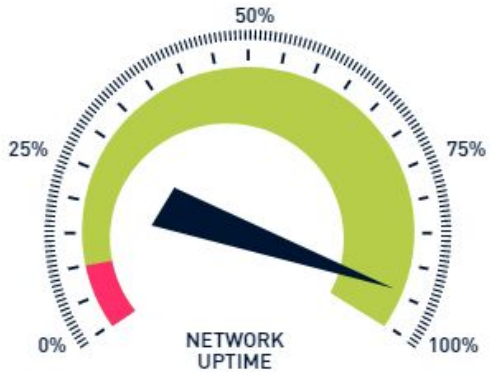
# Resource Management

Capacity and management challenges

The demand for new on-demand technology services and the cost of deploying and managing them.

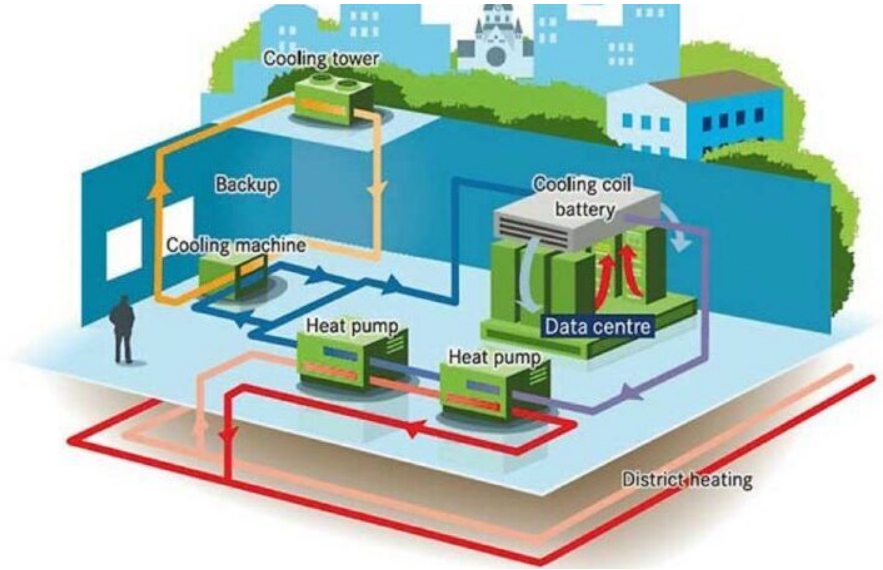
**Challenge 1: Maintaining Availability and Uptime**

If you're using spreadsheets or homegrown tools to manage your server information, you probably already know the information stored can be outdated, inaccurate, or incomplete.  This can prove challenging when unplanned downtime requires troubleshooting, or when attempting to map the power chain.



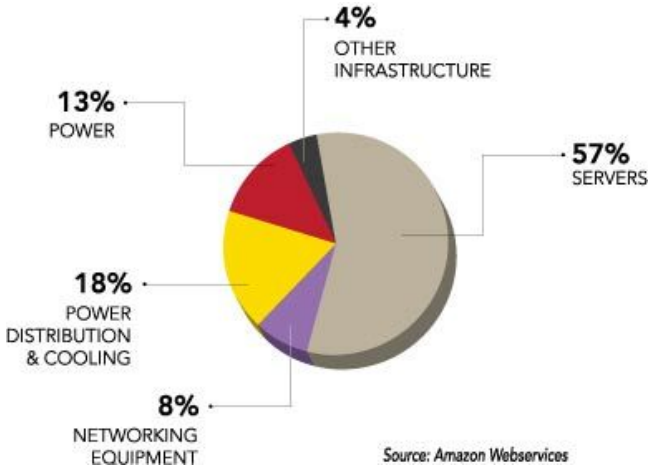
**Challenge 2:  Improving Utilization of Capacity (Power, Cooling, Space)**

In a dynamic data center it is almost impossible to understand how much space, power, and cooling you have; predict when will you run out, which server is the best for a new services, and just how much power is needed to ensure uptime and availability.



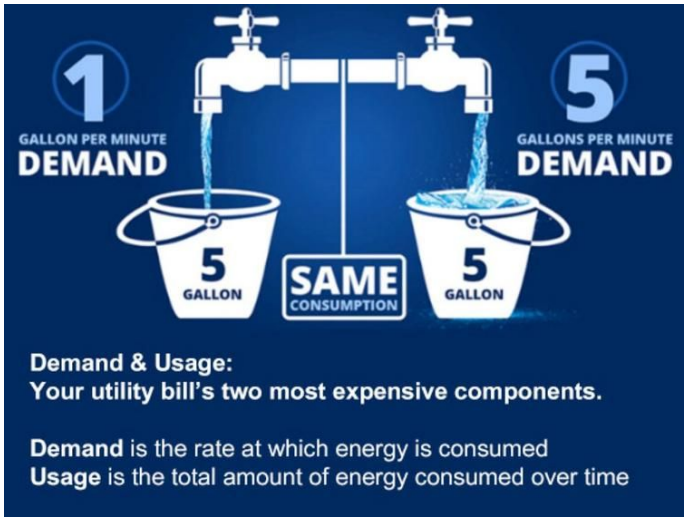
**Challenge 3:  Reporting Reduced Operating Expenses**

It’s not enough to implement solutions that reduce operating expenses, you also have to prove it. According to Uptime institute, “Going forward, enterprise data center managers will need to be able to collect cost and performance data, and articulate their value to the business in order to compete with third party offerings.”



**Challenge 4:  Managing Energy Usage & Costs**

According to a NY Times article, “Most data centers, by design, consume vast amounts of energy in an incongruously wasteful manner…online companies typically run their facilities at maximum capacity around the clock…as a result, data centers can waste 90 percent or more of the electricity they pull off the grid.”



**Challenge 5:  Improving Staff Productivity**

Non-automated or manual systems require facilities and IT staff to spend an extraordinary amount of time logging activities into spreadsheets. This takes away time that can be spent making strategic decisions for the data center and improving service offerings.