Dedicated vs. Cloud:

Comparing dedicated and cloud infrastructure for high availability (HA) and non-high availability applications

Avi Freedman
SCTG Technical Advisor

A White Paper by

SERVERCENTRAL TURINGROUP

Table of Contents

Dedicated vs. cloud
Using auto-scaling to cut costs

One of the most difficult IT decisions facing organizations today is the choice between dedicated servers and cloud solutions. When considering a transition, the largest inhibitors are often the initial investment in cloud infrastructure, and the migration strategy from a dedicated environment to the cloud.

Investing in human and hardware resources

Migrating applications and infrastructure requires a large expenditure of human and financial resources. In a private cloud environment, the initial investment in infrastructure is significant. In addition to the cost of hypervisors, SAN, backup, and network hardware, the human capital required for configurations, maintenance, and training administrative staff is often underestimated.

Finding time

Many dedicated environments cannot afford extended downtime for maintenance, let alone a migration. Finding time amidst day-to-day operations to perform a migration is never easy.

Assessing risk

There is a notable risk in migrating production applications, further increasing cost and complexity. In some environments, it's not possible to revert to the original infrastructure should the migration fail.

This white paper was created to help organizations make the right choice for them, whether it be deploying infrastructure in the cloud or on a dedicated platform. It discusses the many approaches to cloud and dedicated platforms, highlighting a few use cases along the way.

If you have any questions while reading this paper, or would like to discuss any of these topics in more detail, contact us at your convenience:

Avi Freedman
Technical Advisor
Contact us

Hosting terminology

To start our discussion, we'll review terminology and technology definitions used in cloud services to make sure we are all on the same page.

Control fabric:

Provides the orchestration of all cloud services in multiple pools. A control fabric will handle the provisioning of compute, storage, network, and other services operated in the cloud. The control fabric simplifies provisioning and allocating resources in various service pools through the abstraction of discrete resource provisioning amongst resources to an API call for the customer. It also includes high availability by detecting failures of subcomponents and restarting VMs where needed.

High availability:

Refers to systems, system design approaches and associated service implementations that deliver a prearranged level of operational performance to be met during a contractual measurement period.

Hypervisor::

Physical machine host deployed in the cloud that serves as a compute resource and hosts one or more virtual machines.

Over-allocation:

Many service providers oversubscribe the ratio of physical CPU cores to vCPU cores to maximize profit. A common ratio is 5:1—five virtual cores to one physical core. ServerCentral's Enterprise Cloud allocates a 1:1 ratio of physical to virtual CPUs.

Pool:

Group of hardware resources operating in unison to provide a service in the cloud. For example, multiple physical hypervisors may be grouped into a pool to provide a compute pool while SAN and NAS devices may be allocated into a storage pool.

Virtual CPU (vCPU):

Partitions of a CPU. Each CPU can be referred to as a core, but some cloud providers give quotes on vCPUs with multiple cores, whereas many providers use the term core and vCPU interchangeably.

Virtual machine (VM):

A "logical" server with its own CPU, RAM, disk, network, and operating system. VMs live on a hypervisor.

Dedicated hosting

Traditional IT infrastructure deployments are commonly referred to as dedicated hosting and require dedicated hardware with resources allocated to each component. Physical servers and network resources are commonly deployed for each function.

In dedicated hosting, each physical server retains its own set of disks, processors, RAM, and network interfaces. Each component is a discrete physical component and all resources within each component are dedicated to the tasks deployed on the physical machine.

Typically, if a component fails, it can be down for hours or even days. High availability requires duplication of all the resources, which can be very costly and inefficient compared to deploying similar architectures in a cloud environment.

In traditional two-tier application architectures, a dedicated environment would be comprised of two dedicated physical servers, a dedicated switch and bandwidth. The servers are deployed on separate physical machines, each with its own CPU, network, disk, and RAM.

Servers are traditionally over-provisioned to account for the time it takes to order, provision, and deploy new resources. As a result, dedicated

environments are difficult to scale efficiently. Most environments have many resources that are at extremely low utilization but require a full spend on power, operations, and maintenance, and cannot be redeployed to other resources.

In a dedicated environment, resources are typically over-provisioned because it lacks the ability to respond in the event of a resource utilization spike. These resources remain idle and unused while incurring additional cost during normal operation. It is common practice to over-provision a server resource to account for as much as a 50% utilization increase. The ratio is an arbitrary number dependent on each organization's risk mitigation strategies.

Over-allocating resources on every component in a dedicated environment ensures business continuity through a "brute-force" provisioning method of having excess capacity to serve deficiencies elsewhere within the infrastructure.

Cloud hosting

Is cloud hosting really just a mainframe for servers?

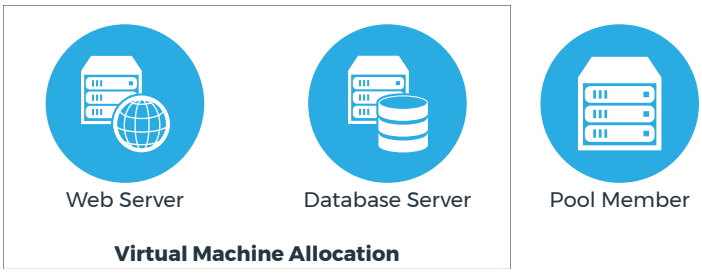
Cloud-based infrastructure is comprised of a pool of machines. The resources in the pool are shared and available for multiple virtual machines (also called instances) to use, and can be reallocated dynamically as needs change. When deployed properly, cloud virtual machines provide high availability, efficient resource usage, higher performance, and with cost predictability.

Software and OS resource management algorithms can affect resource utilization efficiency on a per-application basis. A cloud environment can more efficiently utilize these resources by distributing the load across two virtual machines operating at 80% efficiency compared to a single dedicated server operating at 50% efficiency.

A cloud environment ensures resources that would normally be idle are utilized to provide maximum cost savings. These unused resources normally consume electricity, generate heat (which requires additional cooling), and incur operational wear even while idle.

In a cloud-hosted environment, discrete performance metrics can be maximized by tier or function within the application infrastructure. The ability to redeploy available resources to different tiers contributes to cost containment, thanks to cloud elasticity based on immediate needs.

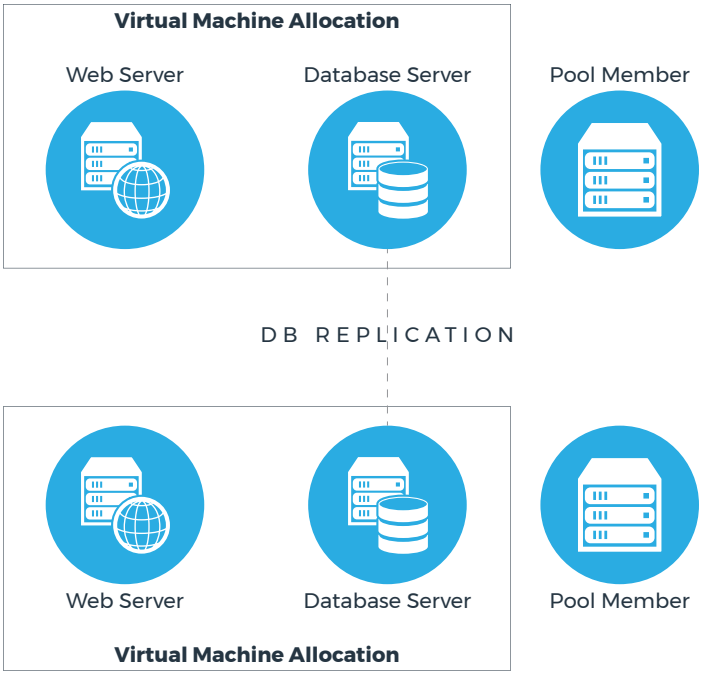
In a cloud-based environment, two-tier application architectures (webserver + database server) can be implemented as follows:



In this example, both servers are virtualized and share the resources of the host machine (hypervisor in a compute-pool) and the servers have preallocated CPU, RAM, disk, and network resources. The virtual machines function as independent logical entities/servers.

Unused resources may be dynamically redeployed to other virtual machines as needs arise. Additional virtual machines may also be hosted on the hypervisor as long as resources are available.

Using two hypervisors that are pool members, a high-availability solution can be deployed as illustrated below. Application and infrastructure availability are increased and resources are more efficiently utilized.



In this scenario, the environment is replicated on two hypervisors ("Pool Members") and data in the application infrastructure is redundant. High availability and increased traffic capacity can be achieved by adding load balancers and using DNS Round Robin or other load balancing approaches (HAProxy, etc.). A fully redundant architecture not only contributes to improved application availability, but when coupled with a load balancer, provides an increase in total system capacity. In an application engineered for the cloud, an increase in available stateless resource handlers can exponentially improve application performance compared to a linear increase in non-cloud environments.

Non-compute resource in the cloud

My servers all have RAID and are not single point of failure. What are the other benefits to centralized storage? Do my file servers already provide this functionality?

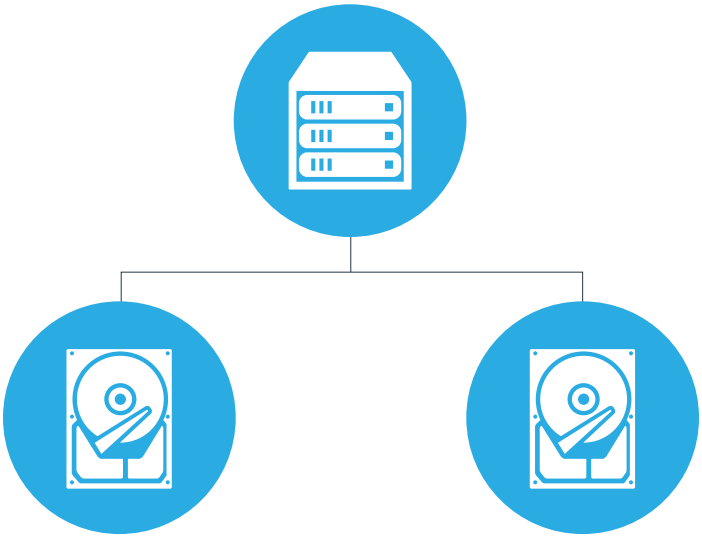
After evaluating dedicated vs. cloud environments, more and more companies are making the informed decision that cloud hosting is a viable alternative to traditional infrastructure. Startups to enterprises are recognizing potential benefits well beyond traditional compute services. For example, traditional resources such as storage and network are more efficiently utilized and become fully redundant when properly deployed in a cloud architecture.

Storage

As resources are shared and allocated among virtual machines, storage infrastructure becomes more flexible. This results in storage benefits that become even more apparent in a fully redundant and highly available storage provisioning scenario.

The ability to replicate, dynamically reprovision, and centralize storage in a cloud environment resolves many of the common issues plaguing IT organizations today. An IT organization with a centralized storage strategy can realize significant savings by reducing the additional physical and human capital resources required to maintain data integrity, availability, and redundancy.

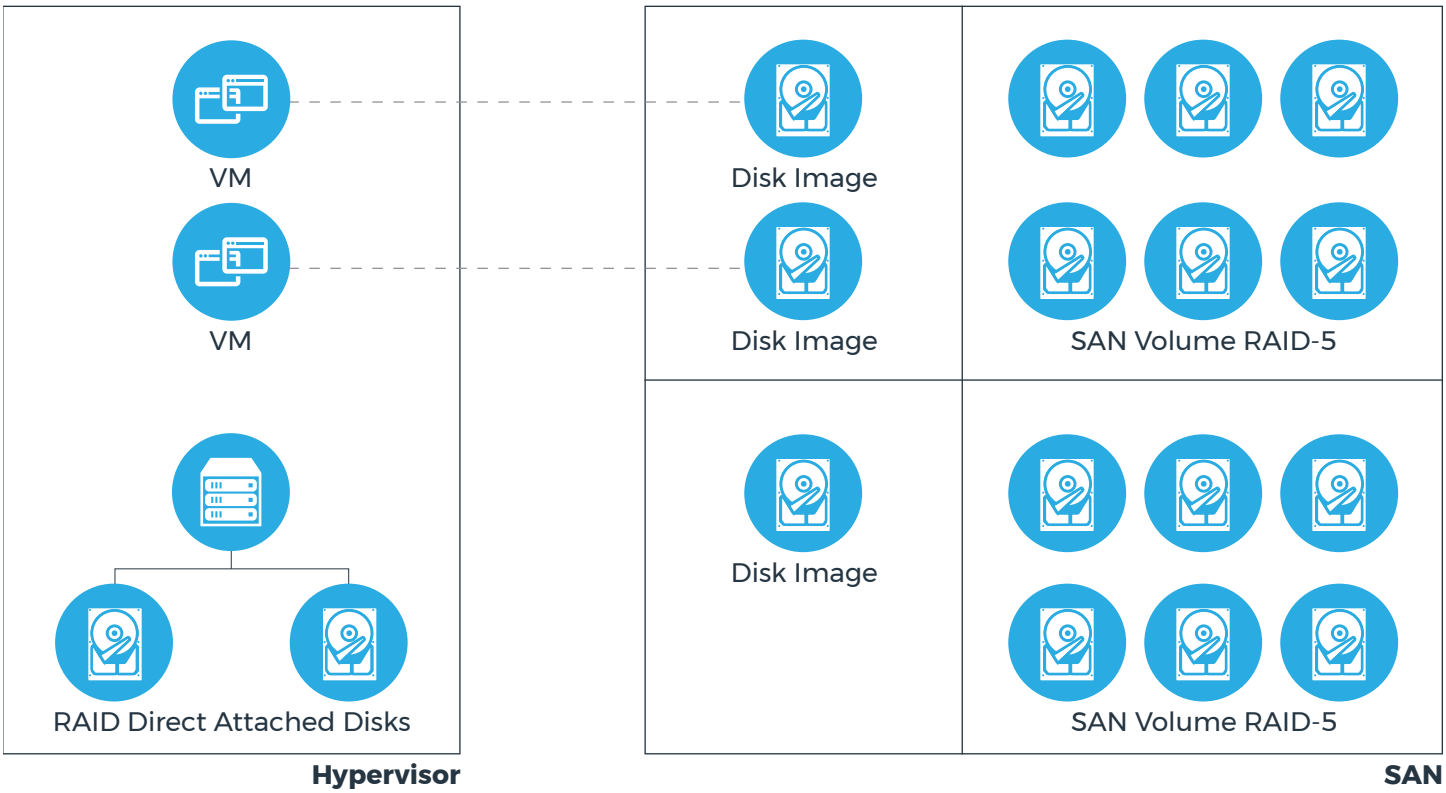
A traditional dedicated server has direct attached storage (SCSI, SATA, SAS) connected to it in a RAID configuration. Common configurations for redundancy are RAID-5, RAID-6, and RAID-10 (for speed). In applications that require high concurrency, low latency and high bandwidth data communications such as a database, Direct Attached Storage (DAS) is preferable.



RAID Direct Attached Disks

Traditional hypervisors host virtual machines that can mount volumes hosted on a SAN using iSCSI or Fibre Channel. Each virtual machine is provided with fully redundant, high performance storage that is centralized.

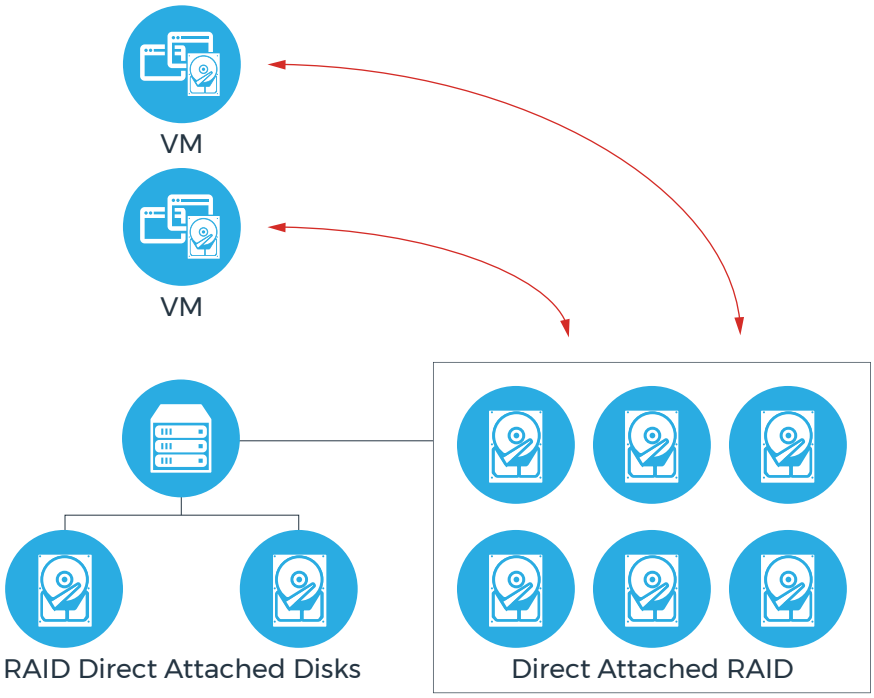
An additional benefit of cloud storage is the cost savings realized through human capital and management related expenses. In today's multi-tiered application environments, centralized storage is a dominant factor in high-performance, cost-efficient computing.



The deployment of centralized storage in IT infrastructure is a primary driver for high performance, highly available computing while providing significant savings.

For example, ServerCentral's high performance Enterprise Cloud hypervisor provides direct-attached RAID arrays on each hypervisor resulting in higher performance and lower latency. This provides the advantages of local disk with the flexibility to restart a virtual machine on a different hypervisor if one fails.

In this scenario, a server is configured with two local disk devices in a RAID-1 (mirrored) configuration for booting the hypervisor and saving state. Disks attached to virtual machines reside on the clustered direct-attached RAID units providing local-attach speed with HA.

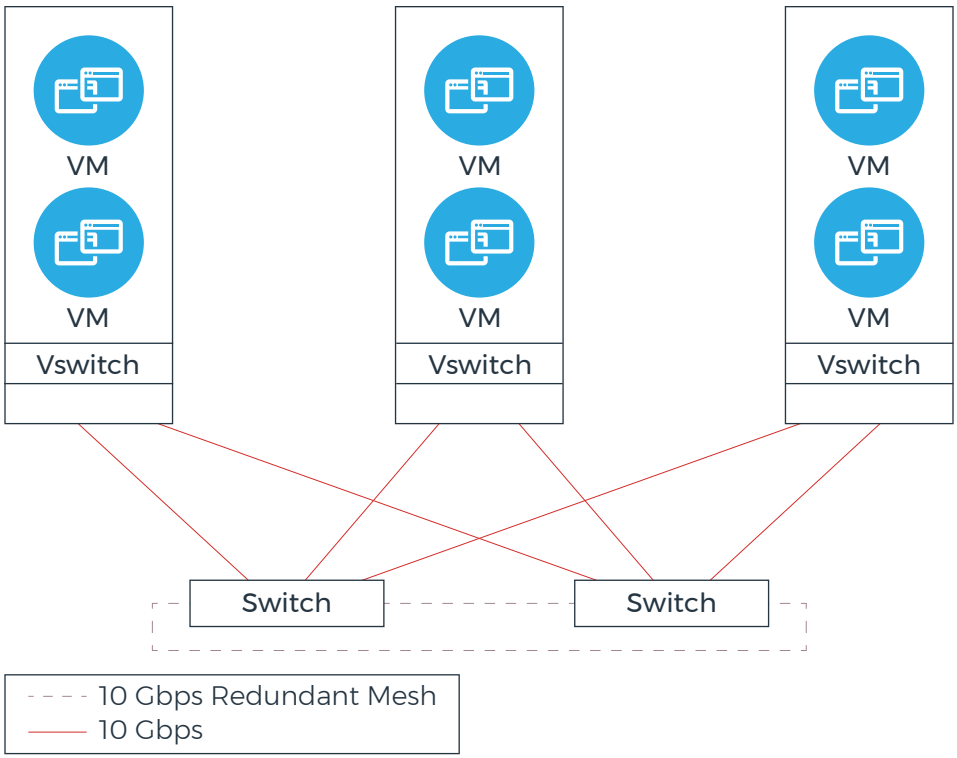


Network

Why do I need cloud networking services when my team has already implemented a fully redundant network?

Like all other resources, network resources are also shared in a cloud environment. Each instance (VM) is allocated a virtual interface and the virtual interfaces are bound to the host adapter interface. In this framework, a host pool can allocate network resources to a virtual machine as they are needed. Hosts can also be dynamically rerouted to a separate network path to mitigate congestion on the network in the event of an increase in traffic.

Additional virtual interfaces can also be created and bound to virtual machines for the purpose of inter-VM communication on the same host without impact to the physical, non-local interface traffic. This reduces latency and bandwidth limitations on the physical network infrastructure.



Cloud services

How will cloud architecture benefit me if I have plenty of capacity today and am having no issues?

Key differentiators in cloud vs. dedicated infrastructure are the services that a well-engineered cloud offering provides. These services aid you in the rapid deployment of new resources such as network, compute, and storage to preserve data integrity through image backups and historical, point-in-time snapshots of the environment.

Services provided by cloud infrastructure include, but are not limited to, the following:

Security

Data security is an ongoing challenge for organizations, and one of the services cloud infrastructure offers is dynamic security architecture and provisioning. Organizations can implement security objects (firewalls, access control templates, and vLAN infrastructure) at any point in the architecture with minimal effort by using the API or Control Fabric Interface.

An administrator can simply "drag and drop" a firewall or network object from the inventory of available components into the infrastructure. This saves time and money by reducing hardware and provisioning (human capital) costs. A network architecture implementation can be reduced from weeks to hours using a Cloud Control Fabric.

Network

Network topologies are an integrated component in a Cloud Control Fabric. The ability to define virtual switches, VLANs, QoS policies, and other topologies can be accomplished through the implementation of network objects in cloud architecture frameworks.

Cloud administration interfaces allow an organization to provision, test, and modify a network topology in minutes with a drag and drop interface.

High availability

Another cloud infrastructure benefit is high availability functionality through the Cloud Control Fabric and the implementation of load balancer as a service. This functionality gives the infrastructure a load-balancer object, which allows for a dynamic deployment anywhere within the application topology. Therefore, customers can "drag and drop" load balancing services anywhere on the implementation, and capacity and availability issues can be rectified in minutes versus hours or days in a dedicated environment.

Data redundancy and availability

Dedicated environments require customization of data retention and availability solutions. Cloud environments provide data retention and availability services through the API or Cloud Control Fabric Interface with an imaging and backup service.

An administrator can make instant API calls to the control fabric that can perform a point-in-time snapshot of a virtual machine, environment, specific disk partition, or infrastructure based on the arguments provided at the time of the request.

These images can be made readily available for deploying new infrastructure or additional VMs within any application tier. In many cases, disaster recovery scenarios in a cloud environment are quickly addressed through these services reducing downtime from days to minutes.

Elasticity and flexibility

Cloud environments allow properly architected application infrastructure to scale as demand increases or decreases. The Cloud Control Fabric API allows an application to scale based on any combination of user defined parameters. When increased traffic load occurs, application, web, or database servers can be created and brought online within minutes to satisfy the resource need. As resource utilization subsides, the extra servers can be removed automatically, which maximizes resource utilization in the customer infrastructure.

Dedicated vs. cloud costs

How do I justify the benefits of dedicated vs. cloud to my CFO?

As with all business decisions, organizations must take into account the financial implications when evaluating dedicated vs. cloud hosting. One difference between the infrastructure deployment models is how accounting is handled. Some organizations prefer to capitalize expenses while others prefer to shift expenses to an operational paradigm.

Human capital resource savings

One of the major factors in deciding between cloud and dedicated infrastructure is cost. While hard costs are an obvious contributing factor, the more subtle soft costs, costs that are unique to and vary significantly by organization, and can be difficult to associate with a number.

Cloud computing shifts the IT burden from your organization to your cloud service provider. In doing so, cloud computing significantly reduces human capital expenses. For example, traditional dedicated environments require every component of an infrastructure to be manually configured and physically available. On the other hand, in cloud environments components are configured in a centralized location from an administrative interface.

Cloud components are treated as objects and any number of the same object may be applied to the infrastructure (as long as it is available in the resource pool). Load balancers, switches, routers, storage, virtual machines, firewalls, and data storage objects are in many cases “drag and drop” or “plan and provision”.

Cloud based infrastructures also offer self-service resource provisioning and management portals for workgroups that have constantly changing demands. This same functionality can be achieved in a dedicated environment, but lacks the dynamic nature of a cloud control panel because of the inability for a dedicated resource control panel to adapt to newly deployed resources in real-time.

Financial considerations

Cloud based infrastructures also offer self-service resource provisioning and management portals for workgroups that have constantly changing demands. This same functionality can be achieved in a dedicated environment, but lacks the dynamic nature of a cloud control panel because of the inability for a dedicated resource control panel to adapt to newly deployed resources in real-time.

Dedicated CapEx vs. OpEx

The Total Cost of Ownership (TCO) in a dedicated environment can be significantly higher than the TCO of a cloud architecture. In dedicated environments, traditional capital expenditures (CapEx) are met with varying operating expenses (OpEx) as well. A challenge that is often overlooked in the OpEx model is the need to maintain components that become defective due to failure. In the OpEx model, these expenses can only be attributed to an individual application or project—applications or projects for which these financial requirements are not consistently assigned.

Cloud CapEx vs. OpEx

In a cloud environment, CapEx is initially realized and the operational expenditure is distributed across the entire infrastructure. As resource utilization increases, infrastructure expansion is classified as an operational expenditure for the entire installation as opposed to capitalization expenditure for a discrete application or project. Savings are realized immediately, lowering overall TCO in the cloud. There is much more to be discussed regarding CapEx vs. OpEx opportunities and challenges. We'll explore this topic further in future papers.

External Decision Factors

As we have discussed throughout this whitepaper, deciding between cloud or dedicated architecture involves serious consideration. Aside from cost, it is important to evaluate other aspects before selecting a solution:

Service Level Agreements (SLAs)

Because a move to the cloud means a loss of some of your control, there should be strong SLAs put into place. Strong SLAs not only protect you in the event of downtime, but demonstrate a service provider's own confidence in the reliability of their cloud. Your cloud provider should offer equal to or better uptime than a data center.

Security

While you may never see the physical hardware running the cloud, it is important to understand how secure that hardware is behind the scenes. Questions should be asked regarding authorized access, safety audits, etc. Better yet, we suggest taking a tour of the facility where the cloud resides to see it firsthand.

Services portfolio and track record

To maximize investment, you should determine a cloud service provider's capabilities to provide additional business services, beyond cloud, that may impact the cost-effectiveness of your operations overall. Is this their core business? Have they been involved in infrastructure for several years? How many customers do they have? What do others say about their service?

Customization flexibility

An out-of-the-box cloud may not meet the unique requirements of your organization. Therefore, you should determine a cloud service provider's ability to deliver you a customized solution that meets your industry or application-specific needs.

Hybrid solutions

An individual cloud solution may not be a one-size-fits-all solution. You should investigate a cloud service provider's ability to support different combinations of cloud and on-premise infrastructures (colo), multi-tenant and private clouds, or some other combination of all of the above.

Trust

Your infrastructure decision will most likely lead to a long-term relationship with your provider. After all, you will be turning over parts or a majority of your data to an outside entity. The right cloud service provider should not only meet your technical requirements, but should demonstrate a history of consultative support, accessibility, and engagement with its customers.

Next steps

Which environment is right for me: Dedicated or Cloud?

The decision to deploy an application or infrastructure in a dedicated or cloud environment involves numerous considerations.

Performance-based arguments for dedicated environments focus on providing superior single-thread performance (CPU), network interface capacity (bandwidth), and localized resource latency. Dedicated environments also increase cost and inefficiency. Cloud deployments may slightly decrease performance. However, cloud deployments provide high-availability, redundancy and human capital and cost savings that many times offset the performance decrease.

Conclusion

"One size fits all" is not an optimal approach—and dedicated vs. cloud is no exception. Each organization differs significantly in their infrastructure needs. These needs will then change dramatically over time as the applications and business scale.

Unique requirements and growth opportunities are leading many organizations to seriously consider hybrid deployments of cloud and dedicated services. For dynamic environments with a high degree of unpredictability in growth and/or seasonality, hybrid architectures will be ideal solutions.

Hybrid solutions fulfill the infrastructure needs for dedicated hardware performance while providing the cost efficiency, reliability, and elasticity of the cloud. Hybrid solutions are truly a best of both worlds opportunity. Talk to your IT leadership and current solution providers to identify the proper requirements for your organization before getting into deep discussions with additional solution providers. ServerCentral welcomes a discussion with you about your entire suite of applications and current infrastructure, and can make recommendations about the best ways to get the flexibility, performance, economics, scalability, and reliability you require—regardless of the solution provider or direction you choose.

About ServerCentral Turing Group

Since 2000, ServerCentral Turing Group has helped enable and transform businesses using technology. Migrations to world-class data centers, transitions to multi-cloud environments, the development of cloud-native applications and the introduction of DevOps inspired business processes have helped grow and transform our clients' businesses making SCTG a trusted advisor and solution provider for companies-big and small.

SERVERCENTRAL
TURINGROUP

111 W. Jackson Blvd.
Suite 1600
Chicago, IL 60604

www.servercentral.com
Toll-Free: (888) 875.4804
Worldwide: +1 (312) 829.1111
sales@servercentral.com