The Elastic Cloud

Now entering its second decade, elastic cloud computing is an essential foundation and driving force of digital transformation. By providing universal access to unlimited amounts of computing resources and storage capacity on a pay-for-what-you-use basis —removing the need for expensive upfront capital outlays—the cloud has democratized IT. It enables organizations of any size to apply AI to data sets of any size. While an increasing number of existing organizations are rapidly shifting large parts of their IT to public cloud platforms, many still resist moving to the cloud. Such resistance is often based on outdated beliefs about inadequate cloud security, availability, and reliability. In fact, since the emergence of the first public cloud offerings more than a decade ago, the rapid evolution and extensive investments of the leading cloud providers have catapulted them to the forefront, outpacing traditional data centers in virtually every measure. This chapter examines in greater detail the rise of cloud computing, its business value, benefits, and risks. It's important that business and governmental leaders understand this paradigm, which fundamentally changes the economics of computing and IT infrastructure. Organizations unable or unwilling to embrace cloud computing will be at a severe disadvantage compared to their competitors. The "elastic cloud" gains its name from the ability to rapidly and dynamically expand and contract to satisfy compute and storage resource needs. This elasticity has transformed software deployment models, the costs of IT, and how capital is allocated. Cloud computing has even transformed entire industries and enabled new ones to emerge. Music, for example, is almost entirely delivered and accessed today through cloud-based services like Spotify and Apple Music, rather than on CDs. Cloud-based streaming media services like Netflix and Amazon Prime Video are rapidly growing, luring viewers away from traditional cable TV and movie theaters. And ride-sharing services like Uber and Lyft would not exist without the cloud. Developers no longer have to invest heavily in hardware to build and deploy a service. They no longer worry about overprovisioning—thus wasting costly resources—or underprovisioning for an application that becomes wildly popular and missing potential customers and revenue. Organizations with large computing operations that can be done in parallel rather than in a linear sequence—such as processing streams of credit card transactions—can get results as quickly as their programs can scale, since using 1,000 servers for one hour costs the same as using one server for 1,000 hours. This elasticity of resources, with no premium for large-scale computation, is unprecedented in the history of IT.

Evolution of the Elastic Cloud:

From Mainframes to Virtualization The evolution of the cloud began with the emergence of mainframe computers back in the 1950s. Mainframes subsequently became the bedrock of enterprise computing for several decades. Whether you booked an airline ticket or withdrew cash at an ATM, you indirectly interacted with a mainframe. Enterprises trusted mainframes to run critical operations because they were designed for "reliability, availability, and serviceability" (RAS), a term popularized by IBM.

John McCarthy, then a professor at MIT, introduced the idea of time-sharing of computing resources on an air defense system in 1959. In 1961, a working demonstration of timesharing called the Compatible Time-Sharing System (CTSS) launched two decades of development across academia and industry. The Multiplexed Information and Computing Service (Multics) and subsequent Unix time-sharing

operating systems propelled academia and businesses to widely adopt time-sharing as a way to access and share expensive, centralized computing resources.

Virtualization—the ability to create private partitions on computational, storage, and networking resources—is the technological innovation that makes today's cloud possible. This is what enables cloud providers like AWS, Azure, and Google Cloud to offer customers private, secure resources while maintaining large pools of servers and storage facilities in central locations, connected by high-speed networks. While the concept emerged in the 1970s, the term virtualization was popularized only in the 1990s. The first system that offered users a virtualized machine was IBM's CP40 operating system, released in the early 1970s. With the CP40, users had their own operating system without having to work with other timesharing users.

Application virtualization was first popularized in the early 1990s with Sun Microsystems' Java. The Java Runtime Environment (JRE) enabled applications to run on any computer that had JRE installed. Until Java, developers had to compile code for each platform their software ran on. This was slow and resource intensive, particularly across the many Unix platforms popular at the time. Java created the ability to run internet-ready applications without having to compile code for each platform. The success of Java ushered in an era of similar tools that supported multiplatform deployment, including Connectix's Virtual PC for Macintosh (1997) 5 and VMware's Workstation (1999).

In the early 2000s, VMware transformed application virtualization by introducing software known as a "hypervisor," which didn't require a host operating system to run. Hypervisor software, along with virtual desktop interfaces (VDIs), made virtualization an easy move for many enterprises at the time. This separation of hardware from the operating system and applications led to the evolution of the enterprise data center, a more flexible version of shared compute and storage resources than mainframes. Major players like HP, VMware, Dell, Oracle, IBM, and others led this market for what we today call the "private cloud."

In parallel, several networking advances led to the development of the public cloud. From 1969, with the initial demonstration of ARPANET (from which the internet was born), 7 to the introduction of the Transmission Control Protocol (TCP), X.25 networks, Internet Protocol (IP), Packet Switching, Frame Relay, Multiprotocol Label Switching (MPLS), and Domain Name System (DNS), both public and private networks advanced quickly. Virtual private network (VPN) technology enabled businesses to use public networks as private networks, which freed them from expensive and slow, dedicated connections between facilities. Commercialized by major telecom giants in the late 1990s and 2000s, VPNs enabled organizations to conduct business online securely. All these advances, plus better and cheaper networking hardware, set the stage for today's public cloud.

The Rise of the Public Cloud

In 2006, Amazon Web Services (AWS) introduced Simple Storage Service (S3) for storing data of any kind and size; Elastic Compute Cloud (EC2), an IaaS virtual machine for computation at any scale; and Simple Queue Service (SQS), for sending and receiving messages between software components of any volume. This marked the introduction of the public cloud, a commercial offering for companies needing to simply and securely collect, store, and analyze their data at any scale. The idea you could "rent" units of compute and storage resources that could be instantly

provisioned and managed by a third party to support operations and users anywhere in the world was revolutionary.

The impact of the public cloud cannot be overstated. It freed enterprises from huge administrative burdens and insulated IT departments from the ongoing cycle of matching resources with demand. Equally important, the public cloud freed IT staff to work with lines of business to better understand requirements and business needs.

Amazon's three initial services effectively paved the way for other cloud service providers to follow. Google entered the cloud business with the launch of the Google App Engine in 2008. 8 Microsoft announced Azure later that year and released its first cloud products in stages from 2009 through early 2010. 9 IBM entered the fray with its acquisition of SoftLayer, the progenitor of IBM Cloud, in 2013.

The range of capabilities available on the public cloud (and consequently, the range of use cases) has increased over the years. AWS, for example, has introduced major advances including the content delivery network CloudFront (2008), Virtual Private Cloud (2009), Relational Database Service (2009), and its serverless offering, Lambda (2014). All these advances, from AWS and others, have created an environment where companies can be born, built, and operated in the cloud. Cloud-born companies can move quickly and scale up or down as needed.

Cloud Computing

Today Cloud computing continues to evolve, being redefined and reimagined with new characteristics, deployment models, and service models. Since 2006, when Google's then-CEO Eric Schmidt popularized the term "cloud" to describe the business model of providing services across the internet, many varying definitions of cloud have led to confusion, skepticism, and market hype. In 2011, the U.S. National Institute of Standards and Technology (NIST) recognized the importance of cloud computing and standardized the definition. NIST defines cloud computing as a "model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."

Today, cloud computing critically underpins and helps drive digital transformation. The mass emergence of digital-native companies today would not be possible without easy, immediate, and affordable access to the scalable computing resources available through the elastic public cloud. Existing organizations are taking note of both the favorable economics and greater agility the cloud offers to enable their own digital transformation efforts, and they are leveraging the growing set of cloud computing features.

Cloud Features

Five core features of cloud computing make it essential to digital transformation:

1. Infinite capacity: Storage and compute resources are essentially unlimited.
2. On-demand self-service: Users can unilaterally provision computing resources without requiring human interaction from the cloud provider
3. Broad network access: Users can operate on the cloud through traditional telecommunication services like Wi-Fi, internet, and mobile (e.g., 3G, 4G, or LTE) on nearly any device, which means cloud computing is accessible anywhere.
4. Resource pooling: Cloud providers serve multiple users through a multitenant model, enabling a pool of physical and virtual resources to be dynamically assigned

and reassigned according to each user's demand— thereby reducing resource costs for all users.
5. Rapid elasticity: Resources can be automatically, seamlessly, and rapidly provisioned and deprovisioned as a user's demand increases or decreases.

Cloud Deployment Models

Beyond its core technical features, two aspects of cloud computing—the deployment model (who owns the infrastructure) and the service model (what type of services are provided)—have significant impacts on business operations. There are three different deployment models, determined by ownership:

- Public cloud is infrastructure available for use by anyone. It is owned, managed, and operated by a business (e.g., AWS, Azure, IBM, or Google Cloud) or a government. The public cloud has gained significant traction with corporations due to its infinite capacity, near-real-time elasticity, strong security, and high reliability.
- Private cloud is infrastructure owned by and operated for the benefit of a single organization—effectively a data center, or collection of data centers, operated on a cloud model by an organization for its exclusive use. An organization's private cloud often has limited elasticity and finite capacity because it is gated by hardware.
- Hybrid cloud combines private and public cloud infrastructures. Hybrid cloud infrastructure is a dynamic space, where public cloud providers are offering up dynamic extensible private cloud environments (e.g., AWS GovCloud) within a public cloud, thereby offering the best of both worlds.

FIGURE 4.1
As business executives define a cloud computing strategy, they must align their digital transformation approach with the pros and cons of the chosen deployment mode.

Cloud Service Models

The second aspect of cloud computing that significantly impacts how organizations leverage the cloud is the service model. The cloud industry has evolved three broad ways in which a business can harness the power of the cloud:

- Infrastructure-as-a-Service (IaaS) comprises the infrastructure building blocks (compute, storage, and networking resources) provisioned and offered on demand. The cloud service provider is responsible for the infrastructure and users have access to their own virtual machines, with control over the operating system, the virtual disk images, IP addresses, etc. AWS Elastic Compute Cloud (EC2), Azure Virtual Machines, IBM Cloud, and Google Compute Engine are the most prominent examples of IaaS.
- Platform-as-a-Service (PaaS) means ready-to-use development platforms that enable users to build, test, and deploy applications on the cloud. The platform manages the underlying infrastructure, operating system, environments, security, availability, scaling, backup, and any associated needs such as a database. AWS Elastic Beanstalk, Azure Web Apps, and Google App Engine are examples of general-purpose cloud platforms.
- Software-as-a-Service (SaaS) refers to software applications hosted on cloud infrastructure (either public or private) and accessed by users through the internet via a web browser. Before the advent of SaaS, businesses typically had to install and run licensed software applications on their own infrastructure and they had to manage

operations like server availability, security, disaster recovery, software patches, and upgrades. Most organizations did not specialize in these hardware and software maintenance capabilities. In the SaaS model, an organization's IT team does not need to worry about those infrastructure details. The SaaS vendor manages it all. In addition, organizations are typically billed annually, usually based on the number of users licensed to access the SaaS offering. SaaS has had a transformational effect on organizations—they can now focus completely on running their business with dramatically lower IT costs and demands. Over the last two decades, several large SaaS businesses such as Salesforce, Workday, ServiceNow, and Slack have emerged. In addition, all traditional software giants like Microsoft, Oracle, SAP, Adobe, and Autodesk now offer SaaS applications.

This rapid evolution of cloud deployment and service models over the last decade has been a major driver and enabler of digital transformation for both existing organizations and startups across industries. As part of their digital transformation initiatives, organizations are migrating their own private cloud or hybrid cloud environments to public clouds. At the same time, we see new businesses born on the elastic public cloud that readily incorporate these new resources as essential components of their digital DNA.

Global Public Cloud Infrastructure

The major cloud service providers—Amazon Web Services, Google Cloud Platform, and Microsoft Azure—compete fiercely for enterprise customers and their workloads. They invest heavily in state-of-the-art hardware (compute, storage, and networking) and software (hypervisors, operating systems, and a broad array of supporting microservices)—all in order to provide the best in connectivity, performance, availability, and scalability. For enterprises, this has helped reduce cloud costs substantially: Microsoft Azure's storage costs have dropped by 98 percent over the last 10 years.

Similarly, the rise in demand and drop in costs have fueled the rapid addition of computing capacity. In 2015, Amazon added enough capacity every day to support a Fortune 500 enterprise. And the pace of cloud growth does not seem to be slowing. In 2018, worldwide public cloud services market revenue grew 21 percent to $175.8 billion, up from $145.3 billion in 2017. It will exceed $278 billion in 2021.

The trend is clear. Intense competition among the public cloud vendors, for whom cloud computing represents a rapidly growing multibillion-dollar line of business, will continue to drive down costs for their customers while increasing the richness of these offerings. The business imperative to embrace cloud computing becomes more compelling each day.

Connectivity to the Cloud

Adoption of cloud services globally is also propelled by improved connectivity of the telecommunications industry. Network speeds around the world are increasing significantly thanks in part to fiber installations in cities and buildings. Average network speed throughout the U.S. is over 18 megabits per second (Mbps)—only tenth in the world. South Korea tops the list at almost 30 Mbps. Worldwide, the average speed is just over 7 Mbps, increasing 15 percent a year.

Historically, fixed networks offered speeds and latencies superior to mobile networks. But continued innovation in mobile network technology—3G and 4G

(third- and fourthgeneration), and Long-Term Evolution (LTE)—has rapidly narrowed the performance gap between fixed and mobile networks. And worldwide demand for next-generation tablets and smartphones pushes carriers to invest in mobile networking infrastructure. Even higher speeds of 5G (fifthgeneration) technology will further accelerate the adoption of cloud computing.

While 5G networks are only in the very early roll-out stages, and estimates about actual speeds abound, it's clear they will be significantly faster than 4G. At the 2018 Consumer Electronics Show in Las Vegas, Qualcomm simulated what 5G speeds would be in San Francisco and Frankfurt. The Frankfurt demo showed download speeds greater than 490 Mbps for a typical user—compared with typical rates of just 20-35 Mbps over today's 4G LTE networks. San Francisco was even faster: 1.4 Gbps (gigabits per second).

The key point for business and governmental leaders is that cloud computing technology and the infrastructure it relies on continue to improve and evolve at a rapid pace. Performance and scalability are getting better all the time—all the more reason to move to the cloud without delay.

Converting CapEx to OpEx

Public cloud IaaS growth of 30-35 percent per year over the last three years 16 illustrates that businesses—particularly those undertaking digital transformations—are moving to the cloud for a variety of technical and financial reasons. As enterprises transition to elastic public clouds, they quickly realize the economic appeal of cloud computing, often described as "converting capital expenses to operating expenses" (CapEx to OpEx) through the pay-as-you-go SaaS, PaaS, and IaaS service models. 17 Rather than tie up capital to buy or license depreciating assets like servers and storage hardware, organizations can instantly access on-demand resources in the cloud of their choice, for which they are billed in a granular fashion based on usage.

Utility-based pricing allows an organization to purchase compute-hours distributed non-uniformly. For example, 100 compute-hours consumed within an eight-hour period, or 100 compute-hours consumed within a two-hour period using quadruple the resources, costs the same. While usage-based bandwidth pricing has long been available in networking, it is a revolutionary concept for compute resources.

The absence of upfront capital expenses, as well as savings in the cost of personnel required to manage and maintain diverse hardware platforms, allows organizations to redirect this freed-up money and invest in their digital transformation efforts, such as adding IoT devices to monitor their supply chain or deploying predictive analytics for better business intelligence.

Additional Benefits of the Elastic Public Cloud

While cost, time, and flexibility advantages are the fundamental reasons to move to the elastic public cloud, there are other important benefits:

• Near-zero maintenance: In the public cloud, businesses no longer need to spend significant resources on software and hardware maintenance, such as operating system upgrades and database indexing. Cloud vendors do these for them.

• Guaranteed availability: In 2017, a major global airline suffered an outage because an employee accidentally turned off the power at its data center. 18 Such unplanned downtime—due to things like operating system upgrade incompatibility, network issues, or server power outages—virtually vanishes in the public cloud. The leading

public cloud providers offer availability guarantees. A 99.99 percent uptime availability, common in the industry, means less than one hour of downtime a year. 19 It is nearly impossible for an in-house IT team to ensure that level of uptime for an enterprise operating globally.
• Cyber and physical security: With the public cloud, organizations benefit from cloud providers' extensive investments in both physical and cyber security managed 24/7 to protect information assets. Public cloud providers continuously install patches for the thousands of vulnerabilities discovered every year and perform penetration testing to identify and fix vulnerabilities. Public cloud providers also offer compliance certification, satisfying local and national security and privacy regulations.
• Latency: Minimizing latency—the lag time between user action and system response—is critical to enabling real-time operations, great customer experiences, and more. The single biggest determinant is the round-trip time between an end user's application (e.g., a web browser) and the infrastructure. Major public cloud providers offer multiple "availability zones"—i.e., physically isolated locations within the same geographic region, connected with low latency, high throughput, and highly redundant networking. For example, AWS spans 53 availability zones in 18 regions globally. With the public cloud, a game developer in Scandinavia, for example, can deploy a mobile application and provide best-in-class latency in every region worldwide without managing a fleet of far-flung data centers.
• Reliable disaster recovery: Today's globally distributed public clouds ensure cross-region replication and the ability to restore to points in time for comprehensive, reliable disaster recovery. For instance, a business whose East Asian data center is impacted by a local political disruption could operate without any interruption from its replica in Australia. Similarly, if files are accidentally destroyed, cloud services allow businesses to restore back to a time when their systems were operating in a normal state. While it is technically possible for any business to set up, manage, and test its own replication and restore services, it would be prohibitively expensive for most organizations.
• Easier and faster development (DevOps): The shift to the cloud enables the new development methodology known as "DevOps" that is gaining widespread popularity and adoption. Software engineers traditionally developed applications on their local workstations but are steadily moving toward developing on the cloud. DevOps combines both software development (Dev) and IT operations (Ops) in much tighter alignment than was previously the case. The cloud gives developers a wider variety of languages and frameworks, up-to-date cloud-based development environments, and easier collaboration and support. With cloud-based containers, engineers can now write code in their preferred development environment that will run reliably in different production environments. All this increases the rate of developing and deploying software for production use
. • Subscription pricing: Cloud computing's utility-based pricing has transitioned software pricing to a subscription model, allowing customers to pay only for their usage. Subscription models for SaaS, PaaS, and IaaS have been popularized in recent years, with pricing typically based on the number of users and compute resources consumed. In most cases, subscription pricing is proportional to different levels of software features selected. This allows businesses to pick and choose what they want, for however long they want, and for any number of users. Even small and medium-sized businesses can optimally access best-in-class software.

• Future-proofing: SaaS allows software producers to rapidly and frequently upgrade products, so customers always have the latest functionality. In the pre-cloud era, businesses often had to wait six months or more between release cycles to get the latest improvements, and rollout could be slow and error-prone. Now, with cloud-based SaaS, businesses continuously receive seamless updates and upgrades, and know they always operate with the newest version.
• Focusing on business, not on IT: In the era of software licenses, businesses had to maintain teams to manage on-premises hosting, software and hardware upgrades, security, performance tuning, and disaster recovery. SaaS offerings free up staff from those tasks, allowing businesses to become nimble and focus on running the business, serving customers, and differentiating from competitors. Computing without Limits The elastic cloud has effectively removed limits on the availability and capacity of computing resources—a fundamental prerequisite to building the new classes of AI and IoT applications that are powering digital transformation. These applications typically deal with massive data sets of terabyte and petabyte scale. Data sets of this size—particularly since they include a wide variety of both structured and unstructured data from numerous sources—present special challenges but are also the essential raw material that makes digital transformation possible. In the next chapter, I turn to the topic of big data in more depth. A Chapter 5 Big Data s computer processing