



Cloud Computing – Storms on the Horizon

Deloitte Center for the Edge

John Hagel, Co-chairman

John Seeley Brown, Independent Co-chairman

Core research team

Blythe Aronowitz

Jitin Asnaani

Indira Gillingham

Sumit Sharma

Interest In cloud computing has been spurred by a confluence of changes in the business and information technology landscape. Today, it is generally viewed as a potentially attractive new form of low cost IT outsourcing, and cloud technology providers and users are focused on tackling the many limitations and challenges of cloud computing, especially in serving enterprise scale needs. Looking ahead, though, we see a series of significant disruptions that will be catalyzed by the evolution of cloud computing.

These disruptions will become progressively more widespread and profound, creating opportunities not only to re-shape the technology industry but all institutional architectures and management practices in an expanding array of other industries. Providers of cloud computing that can provide a compelling shaping view to mobilize other participants will have the potential to carve out a leadership position and reap significant rewards by leveraging their own efforts through the initiatives of many others. As a result, all businesses would be well advised to begin to develop experience with cloud computing platforms at an early stage to better prepare themselves for the disruptions that lie ahead.

The Cloud – what it is, how it got here, and where it is going

Cloud computing is a model for delivering on-demand, self-service computing resources with ubiquitous network access, location-independent resource pooling, rapid elasticity, and a pay per use business model

Rapid experimentation by early cloud providers has created four distinct layers of services: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS), and Business as a Service (BaaS).

- IaaS provides raw utilities such as compute power and electronic storage resources, as services over the network.
- PaaS includes tools and environments to build and operate cloud applications and services;
- SaaS enables on-demand use of software over the internet and private networks;
- BaaS includes application functionality coupled with physical and human resources required to perform a broader set of business activities – typically a major module of activity in a broader business process (e.g., a call center module, as part of the customer service process), or in some cases the complete business process itself (e.g., fully cloud-based supply chain management)

These models of computing are being driven by the confluence of several changes in the business environment and IT landscape. From the business perspective, the trend towards consumer-driven innovation and the growing use of co-opetition and partnership ecosystems is accelerating software development timeframes. Simultaneously, from the IT perspective, several trends focused on increasing the efficiency of software distribution and hardware utilization have converged to enable a cloud computing model, notably early adoption of Software as a Service, proliferation of Hardware Virtualization, and the advent of Utility Computing.

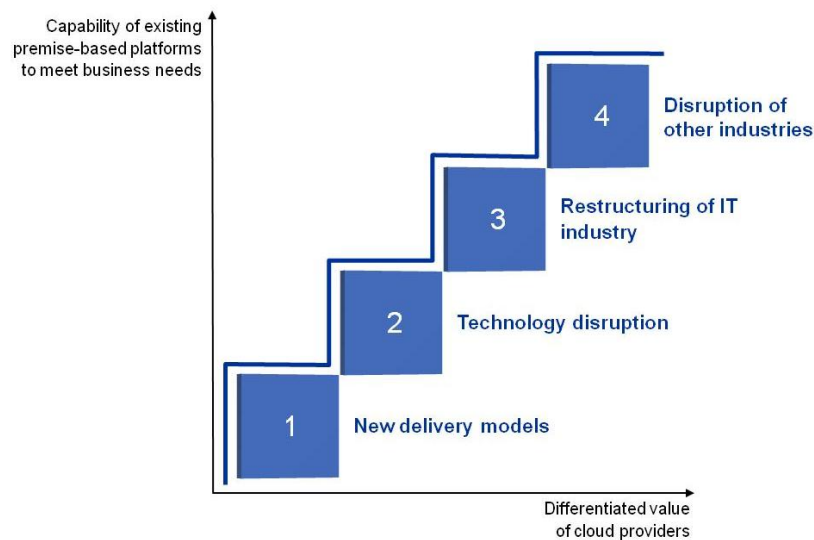
As growing business pressures create the imperative for ever-increasing efficiency, it is no surprise that most current discussions tend to view cloud computing simply as a new, lower cost form of IT outsourcing:

- Lower cost comes from economies of scale including increasing power of virtualization and the ability to move to more commoditized hardware platforms;
- Utility computing helps to turn fixed datacenter costs into more scalable utility costs;
- The power of SaaS – especially in terms of life cycle cost – has enabled rapid software deployment along with easier and faster upgrades.

In combination, these factors have created powerful motivation to drive near-term early adoption in response to growing economic and competitive pressures.

However, while these trends create economics that are very attractive to select companies, they are less compelling to firms which have already made significant investments in premise-based infrastructure. As such, cloud providers are focused on white spaces not currently served well by premise-based solutions. This will lead to significant new technology innovation that will, over time, lead to the emergence of new IT architectures. These new architectures will enable the cloud to become more enterprise-ready, and will compel core IT to move more of the traditional premise-based infrastructure into the cloud.

The adoption of cloud computing will be shaped by a continual iteration of rapidly evolving cloud computing capabilities in areas where existing premise-based infrastructure is not yet able to serve business needs. We see the evolution of cloud computing generating four levels of expanding disruption, driven by a complex interplay of segments of customers with unmet business needs, evolving cloud computing capabilities and new sets of providers emerging to deliver these capabilities to users.



Sequence of disruptions created by the iterative dynamic between cloud users and providers

The four disruptions catalyzed by cloud include 1) the growth of new technology delivery models; 2) the evolution of a new IT architecture to address unmet needs of business ecosystems; 3) rapid adoption leading to the restructuring of the IT industry; and 4) the disruption of other industries beyond high technology based on these new capabilities.

First level of disruption – new delivery models

This first disruption is already occurring on the edges of the core IT industry and builds on the unmet needs of high growth businesses with innovative new delivery models. This disruption has been catalyzed by the global recession and the rapid commoditization of technology.

Target customer segment – While some larger enterprises have become early adopters of cloud computing in limited domains of their IT infrastructure, the bulk of early adoption is in the start-up world, where newly formed companies find the ability to access cloud computing services very attractive, especially in the absence of an existing on-premise infrastructure and affects of the global recession. In fact several features of the recession – including deteriorating business credit facilities, the decimation of Series A venture capital, a growing number of larger cash-rich businesses waiting for the opportune moment to acquire promising new ventures, and so on – has created an environment where low upfront development costs are the “do-or-die” requirement for start-ups.

As an example of this early adoption, as of October 7th 2009, over 1000 SaaS applications had been built on top of cloud services from three leading PaaS and IaaS platforms (Amazon Web Services, Google App Engine, and Force.com). In addition, many SMBs are leveraging cloud services from these three providers to supplement their current premise-based IT architecture.

Besides driving early adoption, cloud startups are also innovating in the way business is conducted in the cloud. By building on each other’s products – directly or through coalitions of interoperable APIs – some startups are enhancing their (joint) value to the marketplace. To illustrate just one example of these new types of interactions, consider The Small Business Web: in early 2009, five software startups, each of which makes a specific leading-edge SaaS application for SMBs, formed the Small Business Web coalition with the purpose of selectively opening up APIs to each other’s products in order to give each of their customer bases access to greater functionality. In so doing, they have not only improved their collective marketing message, but they have in effect created a “federated Salesforce.com”, built with best-in-class “modules” developed by each of these firms. To gain the benefits of scale and scope, they continued to open up the coalition to dozens of other companies selected by a committee consisting of top technology executives from “the Founding Five”. This trend towards interoperability suggests that the cloud has enabled startups to create entirely new business models for both product development and value delivery.

While startups drive the bulk of early adoption, to a lesser degree we also see edges of larger enterprises that are not well supported by central IT (e.g., small and remote branch offices) begin to adopt cloud services. In addition, we are likely to see high growth enterprises and enterprises with highly volatile demand for computing resources begin to move into cloud services to cope with rapid growth and the lack of predictability of the IT load. They will seek to overcome the constraints of IT investment in new premise based infrastructure and the associated lead-times by using cloud providers, especially in consumer oriented, transaction-based web businesses where relatively simple cloud capabilities are required.

Evolution of cloud computing capabilities – Since the early stage companies driving the bulk of early adoption have relatively basic needs, the inability of cloud computing services to replicate sophisticated

requirements in areas like security and SLA enforcement has been less of an obstacle to adoption. Nevertheless, the early adopters at the edge of existing enterprises and the high growth enterprises have started to put pressure on cloud providers to enhance their enterprise level capabilities.

Evolution of cloud computing providers – New entrants such as Amazon and Google have come in from adjacent markets to leverage their deep expertise in managing large and low cost data centers and achieve even greater scale in their core businesses. Other cloud players include SaaS providers operating their own infrastructures to deliver the applications as a service (e.g., Salesforce.com, Intuit and Microsoft). Collectively, these cloud providers are driving the first wave of disruption in the IT industry, focusing on new delivery models including pricing, channels and customer sets. The shift in revenue model for incumbents entails moving from the current economic model, that is characterized by large upfront payments followed by significant ongoing upgrade and support revenues, to one characterized by smaller regular payments based on use. This shift has implications on many areas of the organization including the sales and marketing as the value proposition of the offering changes, and the support organization that will have to manage upgrades, changes by customers, etc. Hardware vendors will continue to see a shift in their customer base as more infrastructure is sourced via large IaaS providers who will have more buying power and volume than individual companies. This first wave of cloud computing adoption is also disrupting traditional channel partnerships as cloud computing makes it easier to reach end users directly and drives the emergence of new channels more specialized in aggregating cloud services for customers. This first level of disruption, however, has been relatively modest at the outset because of relatively limited adoption of cloud computing options and the need to operate at the edge of existing enterprise infrastructures.

Second level of disruption – technology disruption

The second disruption is characterized by the evolution of IT architecture that cloud computing will enable, with the purpose of meeting the needs of a specific and growing set of potential customers - orchestrators who coordinate activity related to end to end extended business processes across a large and diverse network of partners.

Target customer segment – Companies in a variety of demanding global industries such as consumer electronics, motorcycles, and apparel, innovative companies are increasingly adopting an “orchestrator” role – building ever-expanding networks of highly specialized players to deliver customer value. Their needs are not being met by premise-based architectures, so they use manual processes to orchestrate their complex ecosystems. In contrast, leading western firms take the opposite approach: they limit the number of business partners they work with in extended business processes involving supply management, product innovation and distribution channel management in order to reduce complexity, trim costs and extract efficiencies. At least until today, their premise based systems have enabled them to meet customer needs; but as nimbler orchestrators with new cloud-based architectures (as described below) start to emerge, the existing premise based solutions will quickly demonstrate their lack of flexibility and ability to facilitate learning in order to deliver rapid, customer-centric innovation.

In general, orchestrators of complex ecosystems need to coordinate long-lived loosely coupled asynchronous transactions effectively among large numbers of specialized providers. These companies have significant unmet needs and are pioneering management practices to coordinate large and expanding networks but with very limited technology platforms (e.g., phone, fax). The need for

orchestration is growing rapidly as intensifying competition increases the value of large-scale diverse networks in providing flexible, high performance services on a global scale to diverse industries and markets. These emerging “process networks” will enable rapid and reliable innovation through distribution of value-added processes (e.g., manufacturing), as they have started to do in areas as diverse as apparel (Li & Fung), motorcycles assembly (Chongqing), consumer electronics (PortalPlayer and digital camera ODM’s), and even high tech (Cisco Connection Online). (*reference process networks article*)

These target customers, orchestrators, are edge players with limited IT infrastructure in place today and unmet needs that premise based architectures are not able to address.

Evolution of cloud computing capabilities – To understand the capabilities required to serve these customers, let us start by examining their needs, using Rearden Commerce, a major facilitator of corporate travel, as an example. Rearden orchestrates a large ecosystem of diverse participants including corporations, airline companies, car rental companies, etc. To orchestrate such a system, there are key challenges that need to be addressed.

The first challenge is to facilitate the “long-lived”, loosely coupled asynchronous transactions that often accompany multi-party transactions. To illustrate this point, let’s consider the challenge from Rearden’s point of view. On the one hand, the lifecycle of a typical “short-lived” transaction in this industry, such as a one-way flight reservation, is relatively simple: once a customer has purchased the ticket, there are typically only a handful of outcomes, e.g., a successful flight, a change to the flight departure time, a cancellation of the flight altogether, a missed flight, etc. In any of these scenarios, a simple one-step action needs to be taken to compensate for the change. On the other hand, a typical long-lived transaction is far more complex; e.g., the booking of an entire “itinerary” that includes multiple flights, car rentals, hotel stays and restaurant reservations. Such a transaction involves multiple providers and complex interdependencies: any one of the segments of the trip can be altered, and a change to any one segment (e.g., a flight delay) might require a chain reaction of modifications and other actions to compensate for the single alteration. The requisite flexibility and rich exception handling are key properties of long-lived transactions.

Between the beginning and end-point of a complete itinerary, there are numerous different paths that a customer’s ultimate journey may take, despite the initial fixed schedule of flights, car rentals, hotel stays, etc. The “successful” completion of the itinerary is not necessarily the execution of a pre-programmed sequence of actions such as those found in a typical supply chain process – what is known as a **directed graph**. Instead, it is more realistic to assume that the original path will not necessarily be followed and that the only certainties are the beginning and end points of the journey – success will instead depend on fulfilling a number of “constraints”. To concretize this notion, let us consider a fairly simple itinerary, for a corporate traveler such as a senior manager at a consulting firm. In a directed graph world, this senior manager may construct an itinerary that includes a specific flight from Boston to New York, a lunch reservation at Blossom Restaurant in downtown Manhattan, and a specific return flight the same day. On the other hand, a constraint-oriented itinerary would allow the senior manager to specify objectives such as “reach New York before lunch at Blossom”, “meet client for lunch at Blossom at 1pm”, and “return to Boston by 7pm”; the exact mode of transportation chosen to fulfill these objectives will be constrained by policies enforced by the consulting firm for the senior manager role (e.g., only round-trip flights that cost less than \$500 are permissible), and the senior manager can choose specific flights/trains/etc from this constrained list. Ostensibly you get the same result, particularly if all legs of the journey occur as expected. However, notice what happens in each scenario if

the first flight (Boston to New York) is cancelled: in the directed graph, the compensating mechanisms are few – basically the user will have to just book a new flight from scratch, and possibly re-schedule other legs of the journey manually; in the constraint-oriented scenario, this exception can more intelligently be handled – the user can be given a set of options (flights, buses, etc) that are optimized to meet his objective of getting to New York before lunch, and/or it can re-schedule the lunch and subsequent flight if no viable options are available. These benefits are clearly amplified as itineraries become more complex, and this type of non-deterministic, constraint-oriented path from beginning to end is called a **constraint-driven workflow**.

The second challenge that orchestrators face is the need to provide a means for connecting the diverse set of participants to a single platform, including a mechanism for participants to be able to specify and customize their policies – such as the “senior managers may only book round-trip flights that cost less than \$500” policy from above – when needed. Generally, policies are the rules that govern what actions should be taken in certain circumstances, and include IT policies and business policies. For example, a car rental agency might have an IT policy that specifies how much spare compute capacity needs to be available at any time, to deal with spikes in online users; while its business policy might specify what alternative cars will be available to customers if the car they desire is not available at their pick-up location. These types of policies are typically hardcoded, or “embedded”, in the software platform – as such, they are usually static, and accessible only to the platform developer. This is a severe limitation because participants in an ecosystem will need to be able to customize their policies dynamically to account for new innovations (e.g., new flight routes, or availability of new cars) or environmental constraints (e.g., a change in the required check-in time before a flight in the aftermath of 9/11). As such, policies need to be separated from application code and kept in a separate location – called “policy externalization”. This enables the policies of all ecosystem participants to be referenced and mediated by a common platform – in other words, the platform has access to a “federated” repository of ecosystem policies. Such federation is critical to determining how to react to exceptions during the course of a long-lived interaction; for example, what action should be performed to the car rental reservation when the user’s incoming flight is delayed by an hour.

To address these orchestration challenges, a new set of architectural components needs to be developed. The first challenge – the notion of long-lived interactions and constraint-driven workflows – can be overcome through the introduction of two architectural components: **an interaction server** and **an interaction container**.

To understand how these components help, consider the case of a travel itinerary where a delay in flight may interfere with a restaurant appointment – managing this exception may entail automatic changes to the restaurant reservation, or it may require input from the traveler or the restaurant itself to determine the best next step. This need for exception handling is critical to managing long-lived transactions, and is enabled by a robust and explicit interaction container. Analogous to a Java application container, an interaction container manages multiparty interactions by holding a complete “execution context” in which to manage role player interactions and exception; e.g., an interaction container will have access to the full itinerary of the traveler rather than any one isolated segment, enabling it to modify subsequent segments of the journey as exceptions arise. To accomplish this, the container needs to be permeable to allow links to policy extension points (described below) in order to handle exceptions correctly, and also needs to be able to coordinate all the fine-grained services (e.g., car rental service) that comprise the total interaction (i.e., the full itinerary). The interaction container must be scalable to a large number of ecosystem participants, agile enough to support multiple

execution paths based on constraints, and be able to incorporate human Involvement by managing exceptions through a document of record rather than a workflow step.

Of course, there is usually going to be more than one itinerary managed by the orchestrator – this is where the notion of an interaction server becomes critical. In the case of a Rearden-type company, for example, an interaction server might instantiate and manage a hundred interaction containers, each of which manages the complete front-to-back itinerary for a different traveler. Analogous to a J2EE server, the interaction server provides runtime services (e.g., real-time access to fine-grained services such as “booking a car”) to the interaction containers. Critically, it enables a constraint-oriented workflow management engine, as opposed to the typical directed-graph workflow engines, in order to facilitate any one of the infinite paths that any one long-lived transaction might take. In addition, the interaction server supports lifecycle management services, systems and business management services, and services to access and enforce policy. A corollary to the lifecycle management service is that the interaction server must also enable business logic to be linked to infrastructure services so that SLAs can be adequately managed; e.g., if a snowstorm at a major airport were to result in a sudden spike in canceled flights and modified itineraries, the interaction server would dynamically instantiate more resources (e.g., processing power, memory, network bandwidth) in order to deal with the sudden increase in activity.

The combination of these interaction components enable constraint-based workflows with advanced exception handling which cannot be fulfilled by existing technologies, which are primarily built to accommodate linear workflows with minimal exception handling.

To address the second challenge – the need to federate the policies of diverse ecosystem partners – two more architectural components are required: an explicit and distinct **policy engine** that mediates policy differences among participants, and **policy extension points** to enable access to policy in a standardized way.

In the travel itinerary example, each travel provider (e.g., car rental company) needs to be able to publish its business policies (such as contingency actions in case of a delay, cancellation or other scenario) to some sort of federated repository, so that contingency steps can be executed should the scenario constraints be met. The policy engine solves this need by housing these types of constraints in a repository that is external to the interaction server. The repository itself has no logic, but has an engine that interfaces with the repository in real-time. The policy engine must support: federation of policy from multiple participants; versioning of business rules; management of policies which are effective from one point in time to another; dynamically interpreting context to determine the applicability of certain policies; and critically, human Intervention to manage policy constraints and exceptions. The policy engine allows these constraints to be managed and modified in real-time on an as-needed basis by Rearden and its partners, as opposed to typical solutions today where policies are hard-wired into the software and can only be modified by a highly skilled technologist at the IT provider, as per the provider’s software release schedule.

In this travel itinerary scenario, the itineraries are managed by an interaction container; however there also needs to be a mechanism to allow changes to itineraries to be checked in real-time against the constraints in the policy engine in order to enable the correct compensating steps to be taken. This mechanism is enabled by policy extension points, which provide a means for interactions within an interaction container to communicate with the federated policy engine. Policy extension points are enabled by the interaction container, and must be exposed and formally declared. In contrast to

technologies today – which tend to involve rigid, tightly-coupled policies that cannot be easily reconciled – policy extension points allow policies among different participants and across the solution stack to be harmonized in real-time.

In totality, these four new architectural components create a significant evolutionary step in the overall approach to IT system design. We refer to this as an “outside-in approach”, in contrast to today’s traditional “inside-out” approach, which lacks the capability to easily deal with transactions across multiple parties. The four key tenets of this outside-in IT architecture include:

- An ability to connect a very diverse set of external parties to a common platform
- A federated policy model that enables autonomous entities (i.e., ecosystem partners) to be able to set business policies and preferences
- An ability to facilitate coarse-grained, long-lived transactions
- An inherently “pessimistic” view on whether every step of the transaction will be completed in exactly the same manner as initially intended with a much greater emphasis as a result on compensation mechanisms and other approaches to cope with unanticipated developments.

The impact of this evolution is the movement away from workflow oriented models to constraint-based platforms. This enables the orchestrator to accommodate changing circumstances during the life of a transaction to complete the transaction successfully, albeit often in a very different form from the one originally designated. Going back to the itinerary example, the outside-in approach enables Rearden to handle an exception such as a delayed flight simply as another condition to be resolved by the constraint-based policy engine, allowing a number of different compensating actions to be dynamically executed depending on the specific policies of the providers. If Rearden were limited by the traditional inside-out approach, a delayed flight would cause all subsequent legs of the journey to throw exceptions as the transaction would have deemed to have “failed”. The difference in these two approaches becomes profound when you consider the vast number of interactions that orchestrators facilitate, whether in the travel or apparel industries today, or the growing number of other ecosystem-oriented industries of tomorrow. Suffice it to say, the design, implementation and federation of these architectural innovations lend themselves to cloud-based solutions far more readily than to the current premise-based IT infrastructures, due to the inherent “shared services” nature of the cloud.

Evolution of cloud computing providers - The new generation of cloud providers will likely be vertically integrated across infrastructure and cloud management. They will start with a few core applications but will increasingly accommodate third party applications delivered as services. They will actively orchestrate interactions like Rearden Commerce, rather than simply aggregating applications like Salesforce.com. These companies will include tech-savvy orchestrators who have custom developed very sophisticated new IT platforms to coordinate their large and expanding networks. They will provide proof points and early reference models that will inspire a new generation of tech entrepreneurs to design and provide new cloud based architectures to serve themselves and the wave 2 target customers, i.e., the relative low-tech process network companies like Li and Fung.

Already, there are real companies out there that already have made significant progress in implementing these new architectures. We have discussed the example of course Rearden Commerce, an innovator whose solution serves as “personal travel assistant” to 5,600 companies, 160 thousand merchants and 2.8 million users. Another example is TradeCard, which provides a supply chain collaboration platform for 4000+ manufacturers, retailers and their trading partners specifically focused on serving the complex financing needs of various participants in supply chain operations. These

companies have already started building out the outside-in architecture and are influencing customers to adopt the same by enabling, for instance, federated policy engines. Although these companies have started in specific markets like travel and trade financing, they are building a critical mass of participants and are becoming the first wave of cloud providers driving the new architectural innovations.

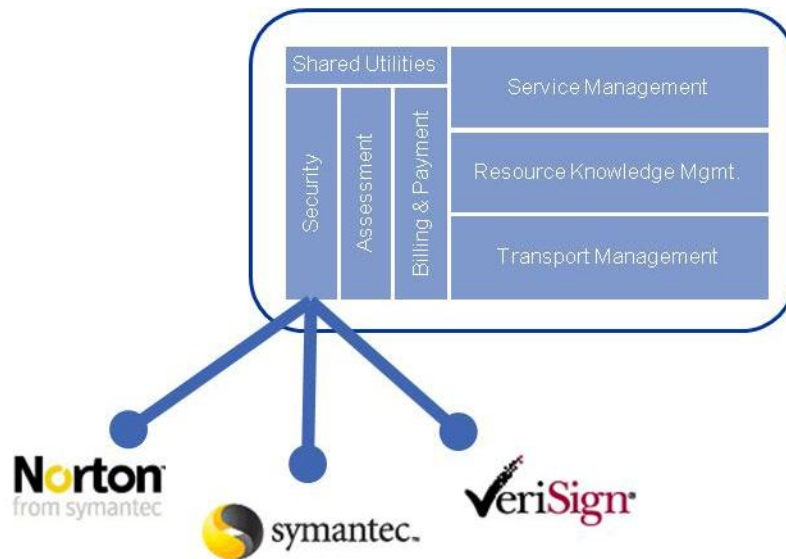
Third level of disruption – restructuring of IT industry

The third disruption will result in the restructuring of the cloud computing industry, driven by rapid adoption of cloud computing services and the resulting pressures placed on providers to deliver best-in-class service at each layer of the stack.

Target customer segments – Having built significant positions at the edge of existing large scale enterprises – both in terms of serving start-ups that are scaling rapidly and orchestrators of large scale business ecosystems – cloud computing providers will now be in a better position to develop the full range of capabilities required to serve the core needs of large scale enterprises. As such, for the first time, the key adopters of cloud computing will be traditional enterprises who finally see a compelling value proposition to move away from premise-based infrastructure. The key elements of the value proposition will include:

- **Compelling economic benefits** of cloud computing: particularly the scalability, ability for low-cost upgrades, and energy efficiency
- **Significant differentiation** emerging from the second wave of disruption in terms of functionality not available from premise based datacenters or private clouds
- **Increasing ability to match and surpass traditional premise based platforms** in terms of basic functionality like security and reliability. This will be facilitated by the availability of robust enabling services and diversity of enterprise-ready applications that will be built with capabilities for handling peak activity, failover, reliability, and other enterprise must-haves

Evolution of cloud computing capabilities – At this point, there will be greater emphasis on development of a full range of capabilities required to serve the core needs of large scale enterprises. To leverage these innovation efforts, cloud providers will begin to more tightly focus on specific layers of the cloud computing stack and, in key layers, develop “**service grids**”, whose purpose is to aggregate atomic services and deliver them to users within guaranteed performance parameters. These service grids will utilize federation frameworks that enable providers to integrate third party services and manage these disparate services on an on-going basis. This will create significant economic leverage by allowing grid providers to free up resources to fund growth rather than having to develop atomic services themselves.



Example of service grid for enabling services, highlighting aggregation of application security services

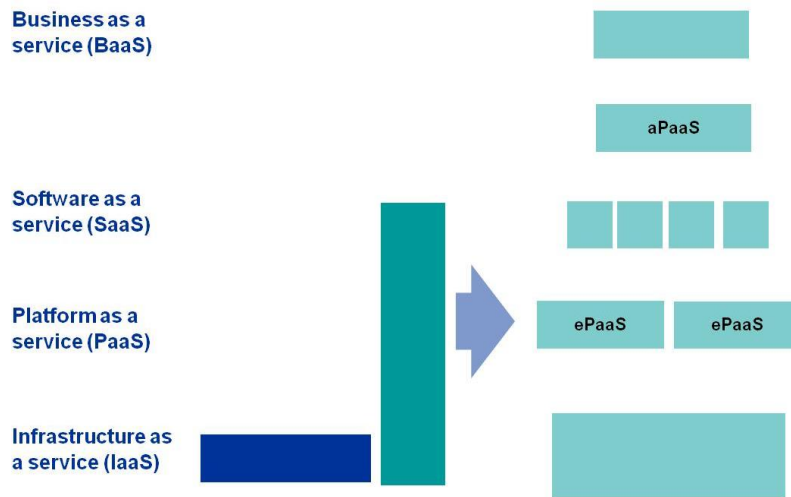
The service grid enhances the cloud by introducing the ability to manage all hosted services according to pre-determined standards; i.e., SLAs. Clouds in general are not formed with registries or other infrastructure necessary to support service composition and governance, whereas service grids inherently have the ability to enforce and harmonize policies across both the business and infrastructural layers within or across the boundaries of the service grid. Policy enforcement is possible because the grid will be able to interact with the externalized policies of third parties. As such, the interactions orchestrated by the grid can be managed by rules relating to business logic, infrastructure provisioning, and regulatory requirements. For example, a service grid can be governed by infrastructure provisioning rules that specify the minimal amount of network, server or storage capacity required to enforce a certain SLA policy. Another example, particularly critical to regulated industries such as financial services or health care, are service grids that can be managed in compliance with industry and domain standards (e.g., ITIL, PCI, SOX, HIPAA) – this compliance is critical for enterprise businesses to achieve a certain comfort level with moving to the cloud, and it also provides the contexts in which lower level services (e.g., security) become especially relevant.

To support enterprise functionality in the cloud ecosystem, the new service grid architectural component will provide managed services including shared utilities, service management, resource knowledge management, and transport management. The service grid also has the ability to manage an integrated SLA based on a bundle of cloud services targeted to a specific business need and in doing so will have advanced business conflict resolution capabilities within and across service grids. This is in contrast to existing architectures that consist of a federation of disparate services that are assembled ad-hoc, with varying levels of service guarantees and policy externalization and hence no integrated SLA enforcement or conflict resolution.

The service grid will help to accelerate innovation at various levels of the cloud computing stack and expand the addressable market for cloud computing within large conventional enterprises, creating a virtuous cycle of innovation leading to broader adoption which in turn funds new waves of innovation.

Evolution of cloud computing providers – As this new generation of cloud users and providers gain critical mass and scale, we will start to see early stage, vertically integrated players in the cloud computing arena begin to unbundle and specialize, followed by consolidation and concentration in key layers as more focused players begin to reap the benefits of economies of scale and scope. These trends will help to accelerate broader adoption of cloud computing platforms serving core enterprise needs.

This is analogous to the disruption to the Personal Computing (PC) industry in the 1970s. Recall that, initially, computing was delivered through vertically integrated providers such as IBM and Univac. As computing gained traction and users demanded higher performance for specific components, leading providers such as IBM restructured the stack – disintegrating their vertically integrated stack and enabling themselves and other players to create best-in-class functionality at specific layers, such as the CPU and operating system. Similarly, cloud computing providers will be pressured by users to provide best-in-class functionality at each level, and so the industry will restructure as in the diagram below.



Restructuring of IT Industry created by the third level of disruption

The existing industry verticals in IaaS and SaaS will restructure into five distinct layers/arenas of cloud computing that are likely to be driven by focused and specialized players:

- **Infrastructure as a service (IaaS) providers** – These players will focus on building and operating large scale data centers providing sophisticated infrastructure management services to optimize utilization of capital intensive computing, storage and network facilities
- **Enabling platform as a service (ePaaS) providers** – These players will focus on managing service grids that source and aggregate enabling services like security, performance management and data translation. In the ePaaS layer, the services aggregated by the service grid will be largely transparent to end users but critical to the application developers building application services at the next layer. These service grids may be provided by specialized independent businesses or by large user enterprises who offer their enabling services to other enterprises. The service grids

will be targeted by domain of expertise; e.g., application security services; or SOX compliance services for financial institutions.

- **Specialized software as a service (SaaS) providers** – These will be highly specialized developers of enabling and application services that will leverage ePaaS platforms described above. These providers will also include a growing number of “user” enterprises who discover the benefits of “exposing” key elements of their business operations as services to be consumed by other enterprises.
- **Application platform as a service (aPaaS) providers** – These players will focus on managing service grids that source and aggregate application services. These players will specialize in particular application domains, whether defined horizontally (e.g., human resource management, customer relationship management), or defined vertically (e.g., financial services, health care). Their focus will be on providing aggregation platforms for a vast array of more specialized application service providers, offering specialized services like SLA management and service directories, enhanced by deep domain expertise to help users configure the appropriate bundles of application services. A critical role of these aPaaS providers will be to enable cloud users to create new coarse-grained business services, composed of granular services available through the aPaaS platform. For example, a financial services aPaaS might enable a financial institution to construct a new loan product by aggregating atomic services such as identity verification, credit history checking, credit risk modeling, etc. As a result, through the aPaaS, the financial institution is able to easily construct a new innovative coarse-grained product by piecing together several best-in-class atomic services which it would otherwise need to create or source through in-house resources.
- **Business as a Service (BaaS) providers** – These will be organizations that integrate application functionality with physical and human resources required to perform a broader set of business activities – typically a major module of activity in a broader business process or in some cases the complete business process itself. One early example of a BaaS provider is Amazon’s logistics offering, which includes a platform plus physical warehouse and distribution facilities. Another is LiveOps’ managed call center services, which includes a software platform along with human resources (i.e., call center operators). More potential BaaS services will be created as a result of the fourth wave of disruption in which other industries harness the capabilities of the cloud.

In addition to these five layers, **specialized professional service firms** will also play crucial roles in cloud adoption and usage, by organizing around specific domains to help end users determine how to most effectively leverage the services of cloud computing providers in their business operations. They will help client organizations to adopt the right mix of services and applications. This could lead to a significant and perhaps disruptive shift from focus on technology design and integration to deep understand of business context and economic drivers to help clients get maximum value from these platforms.

The integrative layers of the evolving cloud computing industry – i.e., the IaaS, ePaaS, and aPaaS layers that will focus on aggregating the components at lower levels of the stack – are likely to become highly concentrated and consolidated. Because they dis-intermediate and commoditize the layers below them, these layers will become the key “control points” for the industry, i.e., these are the lucrative roles that cloud computing leaders will dominate. The two lower levels, IaaS and ePaaS, will serve as the control points for the IT Providers, while aPaaS will be a control point for leading players in several diverse industries and functions.

In contrast to these concentrated control points, the SaaS layer is likely to see a high degree of fragmentation as more specialized players find ways to leverage the resources of the ePaaS and IaaS layers below it. Finally, the BaaS layer may or may not become fragmented within particular domains, depending on economies of scale and scope in the broader business activities; for example, fulfillment is likely to become very concentrated due to the network effects and the importance of economies of scale in that business.

As this re-shaping of the cloud computing industry evolves, it is likely to put greater and greater pressure on existing leaders of the IT industry. A broader array of enterprise level IT infrastructure and applications will become addressable through this more-specialized and scale-driven cloud computing industry and traditional premise-based IT solutions will retreat to narrower niches. As a result, existing leaders of the IT industry will need to find ways to carve out leading roles in the key control points or risk being pushed into narrower and narrower niche roles in other layers of the cloud computing industry or in a shrinking “pre-cloud” arena. New players in the IT industry, either existing scale companies from adjacent arenas like e-commerce and search or completely new entrants riding the cloud computing disruption, will emerge as leaders in the IT industry, displacing many of the traditional leaders. As a result of these disruptive developments, the IT industry is likely to be significantly transformed, both in terms of concentration/fragmentation trends and in terms of the identities of the leaders of the industry.

Who are the likely leaders at each layer of the cloud computing stack?

- Infrastructure as a service providers are likely to come from adjacent arenas where they can leverage the scale they are building in data center operations today – e.g., Amazon and Google
- Enabling platform as a service providers are most likely to be led by new entrants pioneering the architectural innovations required to support distributed ecosystems of end users, although some existing players like Microsoft and HP are potential candidates to play leading roles in this arena
- Application platform as a service providers are most likely to be led by either new entrants focused explicitly on this layer of opportunity, leveraging the resources of the lower layers, or by early entrants into the application as a service arena like Salesforce.com who rapidly move from a product to a platform focus
- SaaS, BaaS, and the accompanying Professional Services are likely to be more fragmented, with concentration possibly emerging around specific industry or functional domains.

Fourth level of disruption – spreading disruptions to non-IT industries

As players across the business landscape increasingly collaborate via cloud-based services, the evolving capabilities of cloud computing will catalyze significant disruptions to a broader and broader array of industries. This will be driven by companies that figure out ways to challenge industry incumbents by leveraging cloud computing to provide significantly more value at lower cost to customers in these other industries.

Health Care Industry – One prime example of this disruption will be in the health care industry, where patient records and shared utility services will be shaped by savvy cloud players. The rising cost for health care and the push for reform from both consumers and government have positioned the industry to benefit greatly from cloud. Savvy technology players will drive change in health care by using cloud computing as a platform to deliver more value to consumers at a lower cost.

Previous attempts at transformation of the health care industry, primarily through Electronic Medical Records (or EMRs) have been mixed, at best. Previous efforts were characterized by demand driven primarily by providers, significant costs to implement, low perceived ROI, significant change management issues across constituent groups, and technology challenges such as painful enterprise integration and security/privacy management. However, while EMR adoption rate is still slow, consumer demand for self managing capability will drive the development of patient health record (PHR) management. These efforts – characterized by demand driven by proactive consumers, wellness and chronic illness management, the need to share management of care for aging parents/relatives, improving patient safety, and aggregation of valuable data for research – will be considerably aided by the cloud, which overcomes the host of technology and coordination issues. In such a system, the key tenets of personal health record management include:

- Control and ownership resides with consumer
- Access to a health record granted based on designation
- Compliance with regulatory and privacy requirements
- Collaboration enabled among multiple care teams
- Interoperability across ecosystem of providers and consumers

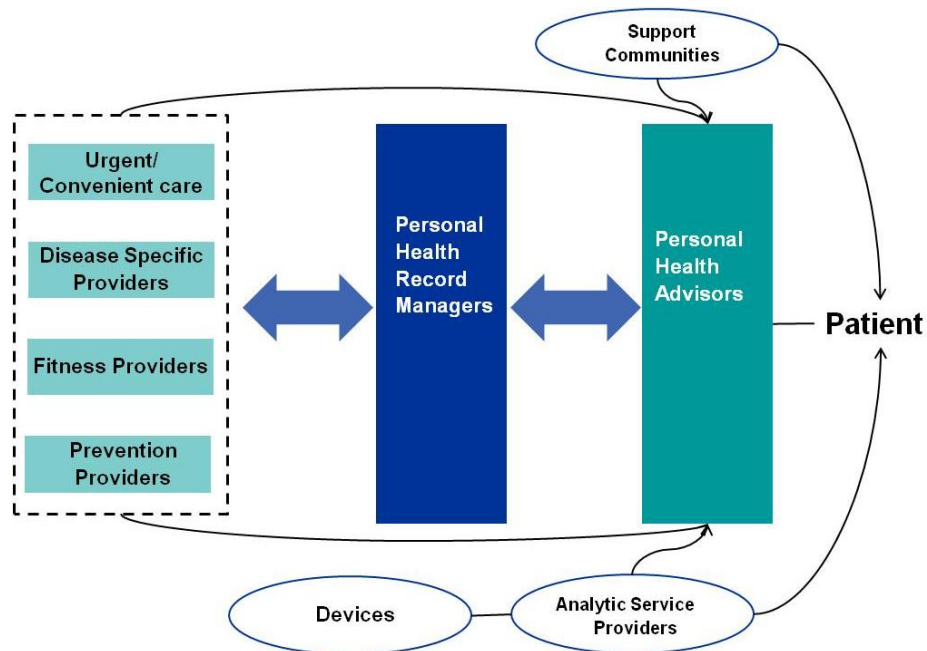
A health care service grid – with all the inherent attributes of a cloud-based solution such as lower costs, ease of collaboration across participants, open standards, higher performance, ease of use – will enable the industry to shift to this information-driven paradigm where previous technologies have not. Indeed, health care is likely to be one of the first industries disrupted, thanks in large part to the current economic environment: sensitivity from the economic crisis leading to concern around finding affordable care, and the proactive attempt to manage health care expenses while improving the quality of care.

Cloud platform adoption will be driven from the edge to the core. At the edge, adoption of personal health records is being driven first by the chronically ill to manage multiple treatment plans, medications, doctor visits, and the vast amount of information and data provided on their ailments. This segment is finding comfort in having the ability to share information within a community, either with consumers diagnosed with the same illnesses, or exchanging in meaningful interactions with experts in specific health domains. The next segment of consumers likely to adopt PHR's will be in the wellness segment; i.e., those who are proactive in maintaining or improving targets to achieve peak health. One example of this is amateur athletes or fitness "fanatics" who use a range of data from certain bio

markers, nutritional data, physical activity, and so on, to analyze in order to make improvements to their fitness regime. Both sets of edge consumers benefit from the ability to manage and share their own data, and are motivated to use PHRs as a tool for managing their health.

As early adoption gains traction, the benefits of PHR management will become more compelling to core health care providers, and patient data will be aggregated, analyzed and shared with industry participants to increase patient outcomes and treatment efficacy; at the same time, PHR providers will continue to create enriched record management features for patients to manage, personalize, and share their records. As this disruption unfolds, an increase in personal health record adoption will spur a renewed interest in implementing electronic medical records. The convergence of PHR and EMR will create comprehensive patient records, including integration of data from multiple repositories/ sources.

Over time, the drive toward lowering health care costs through preventative care will result in a reshaping of the health care industry and the emergence of two new types of providers –personal health advisors and personal health managers. Personal health advisors will provide highly specialized services to consumers by building deep and lasting relationships with the consumer. They will use aggregated data from multiple sources to interact with and recommend actions to proactively maintain health. Another class of providers, personal health record managers, will take on the role of aggregating data from disparate sources such as insurance companies, care providers, ancillary services. The integration of data is invaluable to all participants in the care continuum as it provides a comprehensive history of a consumer’s health, allowing for robust analytics of this data and better care based on more information being available to care providers.



Potential new structure of the health care industry enabled by the fourth level of disruption

Ultimately, the industry will move towards universal access, where data is shared across geographical boundaries for improved patient care everywhere.

Other Industries – Other industries that are likely to be disrupted include financial services, energy, and media. In financial services, the emergence of integrated personal financial management grids will enable unified management of diverse and disparate financial accounts by wealth managers, advisors and consumers who seek to optimize portfolio management across accounts. A financial utilities grid will also enable universal access to commoditized processes such as check processing and financial transaction processing.

In the energy industry, smart grids and power management systems will increase connectivity, automation, and coordination between electricity suppliers, consumers, and networks, while cap and trade platforms will enable universal energy credit trading. The media industry will benefit from digital content service grids, which will enable access to massive quantities of digital media, customized based on specific customer needs and profiles.

Health care, financial services, energy and media are examples of the likely first initial industries that will be disrupted as cloud computing functionality enters maturity, and will lead to a bowling pin reaction of other reactions across industries and applications in an increasingly inter-connected business environment. In addition, these disruptions will in turn feed the further development of new technology features available on the cloud, magnifying the disruptive power of the cloud across industries.

Implications for talent management

There is a hidden value due to the evolution of cloud computing, which will amplify the four disruptions: faster learning and talent development. Indeed, rapidly growing economic pressures on a global scale put greater emphasis on the ability to access and develop talent.

The impact on talent development will manifest itself in each wave of disruption as the increasing intensity of competition will force companies to look for new ways to create and maintain an advantage in the market. Today, in wave 1, there is already tangible evidence of participants benefiting from greater access to resources on-demand through the cloud. This has enabled rapid learning by shortening experimentation lead times, enabling multiple experiments to be conducted in parallel, creating access to scarce and expensive resources, and facilitating collaboration between cloud participants. For example, companies like Varian corporation run intensive remote Monte Carlo simulations of future product designs on the cloud, leading to more rapid feedback cycles; while consortia like The Small Business Web are creating a framework for stitching together complementary products created by individual service providers, so as to create products whose value is “greater than the sum of its parts”. These types of institutions are breaking the ground on rapid learning and collaborative learning that puts pressure on competing and complementary service providers to do the same.

Looking forward to wave 2, we will see orchestrators building explicit mechanisms and platforms for scalable learning. These players will develop and tap into talent outside the organization through platforms that provide real time feedback. An early example of this is LiveOps, which provides a platform for a cloud-based contact center through which each contact center representative (usually a remote, home-based professional) has access to a customized dashboard with real time performance feedback based on goal-specific metrics. This enables representatives to make rapid, autonomous performance improvements in order to better achieve their metrics. As another example, consider Li & Fung, which has developed a low-tech process for delivering real time performance updates, coaching and benchmarking to its ecosystem of 10,000+ partners in order to facilitate quality improvements across the entire ecosystem.

More and more orchestrators like LiveOps and Li & Fung will recognize the erosion of the traditional model of “scalable efficiency” as a competitive advantage, and adopt a model of “scalable learning” to become the market leaders – a trend that we think other industries will follow as cloud adopters start to tap into the talent management opportunities that have so far been ignored by current discussions about the cloud.

The shaping opportunity

This expanding potential for disruption suggests a significant shaping opportunity in the IT provider landscape, driven by a compelling shaping view of the emerging cloud computing arena and its potential impact on a growing array of industries.

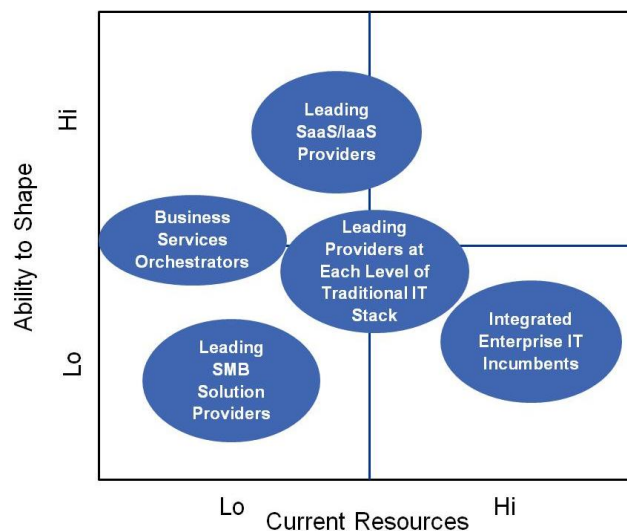
Successful shaping strategies contain several essential elements, developed through in-depth analysis of successful shapers in the past. The first element is a clear and outspoken Shaping View that provides focus and motivation for industry participants by painting a picture of the industry direction and the role

of ecosystem participants. Second, a Shaping Platform that ecosystem participants can leverage to create and capture economic benefits. Third, a series of clear Acts and Assets that demonstrate the would-be Shaper’s conviction in the Shaping View, and builds credibility in the Shaper’s stated role within the View. Finally, a critical mass of ecosystem Participants, which enables increasing returns to scale as more participants engage with the Shaping Platform.

Given our premise that cloud computing will be far more disruptive than most people anticipate, we suggest the following as one plausible Shaping View: “Cloud computing will significantly **accelerate** the movement toward **scalable business ecosystems** focused on **talent development** ... by serving as a catalyst for fundamentally different IT architectures.”

Given this view, a player in the ePaaS layer, one of the primary control points for the IT industry, will likely be in the best position in the cloud stack to create a shaping strategy and platform. A player in the ePaaS layer is naturally positioned to create a platform that reduces the investment required by other services providers; creates protocols around interoperability between services; provides opportunity for a near-infinite number of services to sit on the platform, and can continue innovating on the platform to increase ease of introduction of new services, thus attract an increasing number of customers whose unmet needs can be addressed.

Given the elements needed for a successful shaping strategy, our external and preliminary perspective of current IT Providers suggests how leading providers currently “stack up” with respect to the ability to be Shapers. The horizontal axis summarizes the current assets and resources that could be leveraged for a shaping strategy for the cloud, including the strength and relevance of the customer base, current products and partnerships, and influence on IT architecture. The Y axis portrays a company’s ability to truly shape the industry, by considering factors such as: current commitment to cloud computing, the extent to which leadership is outspoken and regarded as “visionary”, the risk profile of the company, and the culture and agility as pertains to innovation and shaping.



External perspective of the potential for IT providers to become Shapers

Of course, in an ecosystem with increasing returns, the Shaper is not the only player to gain economic benefits – Participants garner value from a shaping strategy as well. Broadly speaking, there are three

types of Participants who can gain from the Shaping Strategy: Influencers, who commit early and prominently to one shaping strategy; Hedgers, who develop products of services to support multiple shaping platforms; and Disciples, who commit exclusively to one shaping platform. (*reference shaping strategies article*) These opportunities apply to all participants in a Shaper ecosystem – the key to creating value is to have clarity and focus around the specific role chosen, i.e., the role that the IT Provider is best positioned to play and the role that it *wants* to play.

Providers who want to be shapers can consider a number of early moves that they might take to improve their likelihood of success. In the future cloud computing ecosystem, one likely play is to target the unmet needs of actual or aspiring orchestrators of specific business ecosystems as identified in the second disruption wave. A shaper can develop a minimal platform as a service to address these unmet needs, riding upon someone else’s infrastructure as a service and then aggressively recruit enabling SaaS (e.g., security, transport) and application SaaS and create ways for these third-parties to connect with clients and other services. Over time, a would-be Shaper can carve out a leadership position in the enabling platform as a service layer of the evolving cloud computing industry.

Bottom line implications for clients

Cloud computing will be far more disruptive than currently anticipated. This creates significant opportunity for new forms of strategic advantage both on the IT Provider side and the “User Enterprise” side, which heightens the need to engage early to build capability and to aggressively pursue the disruptive potential of cloud computing.

Contact us

For further information, please contact:

John Hagel
Co-chairman, Deloitte LLP Center for the Edge
Director, Deloitte Consulting LLP
+1 408 704 2778
jhagel@deloitte.com

Glen Dong
Chief of Staff, Deloitte LLP Center for the Edge
Director, Deloitte Services LP
+1 408 704 4434
gdong@deloitte.com

Christine Brodeur
National Marketing Lead, Deloitte LLP Center for the Edge
Senior Manager, Deloitte Services LP
+1 213 688 4759
cbrodeur@deloitte.com