cloud computing and application performance management

the trust paradigm

Richard Fisher, Managing Consultant, Orange Business Services

Business Services

orange™

contents

introduction

What most affects your IT services experience? Is it availability, ease of use, reliability, predictability, responsiveness or support? Or do you think about the number of CPU minutes and the amount of WAN bandwidth the IT department has purchased?

It is likely that you have experienced the frustration of waiting for information to load on a computer screen when your IT systems were running slowly. Now multiply those emotions and their impact on your productivity by the number of people in your organization, and you start to appreciate the overall business impact of poorly performing applications.

Consider the case in which a company shares the responsibility of delivering applications with a service provider while still being accountable to its end business users and/or customers. Sharing this level of responsibility requires a greater degree of trust than a traditional service model, especially with regard to data integrity and confidentiality as well as service availability and performance. So the question is, how can this trust be established and maintained?

The most common paradigm is that when an IT organization adopts a cloud computing service, they expect to absolve themselves of most (if not all) responsibilities. What they can't see, they can't affect. However, when there is a problem, the CIO and end users contact the IT organization and not the service provider – in the eyes of the business, the members of the IT department are still responsible, whether they like it or not.

This paper examines the need for end-to-end transparency within the application delivery chain for the service provider to earn trust and effectively share the responsibility of managing the performance of a cloud computing-based enterprise architecture. Further elaboration will consider the need for business-critical cloud computing services to be viewed as strategic partnerships and how this trust paradigm is vitally underpinned by managing the end users' experience.

the problem

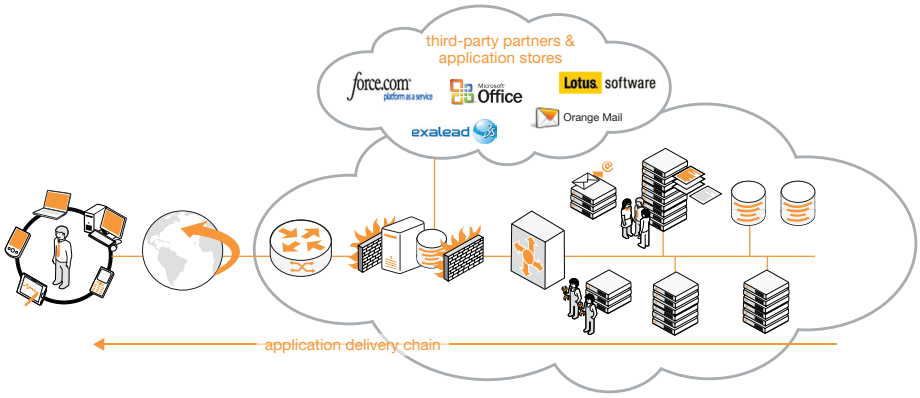


figure 1: the end-to-end application delivery chain is formed of data center components (servers, application code, storage, power, HVAC), third-party partners (private application stores), the network connection to the client (WAN, LAN, Internet, etc.) and the client device (smartphone OS, PC browsers, etc.)

First, we must consider the technical and organizational problems that the service provider and subscriber face when considering performance management in a cloud computing model. *Scope: applicable to all cloud computing models*.

managing complex technology as a service

The centralization of applications to an externally hosted cloud computing service will have an impact on the application response time for some users and the ability of the operational support teams to isolate and resolve service issues.

Operational support groups are typically compartmentalized and specialize on each component of the application delivery chain. The cloud computing model adds to the complexity by introducing an external organization as another layer of specialization. This organizational separation between support groups, combined with the added complexity of virtualized technology, makes it difficult to identify the root cause of application performance issues. We must also consider the potential increase in application response time due to an increase in latency and the number of network boundaries between the end user and the server. Combined, we realize the potential adverse effect that the end user will experience: *the potential for inconsistent or slow application wait times coupled with slow resolution of problems when they arise*.

managing application certification

With the proliferation of consumerization, enterprises will see countless variants of end-user devices, operating systems and browsers accessing cloud-based applications. It is not feasible to test all of these connotations – only a few variants could be certified, and the remaining would be subject to the availability of the support teams to resolve issues. To enable the flexibility that consumerization demands, applications will no longer be certified for use on client devices, but rather problems will be handled dynamically and reactively. *This will require the means to detect, quickly isolate and find the root cause of performance problems at a much faster rate than operations groups are currently used to*.

Perhaps less obvious are the psychological implications of sharing performance management responsibilities. *Scope: applicable to all cloud computing models*.

managing customer expectations

Migrating to the cloud carries the expectation that the new service will be an improvement to or at least deliver the same quality as that provided by the legacy architecture. After all, it is only human nature to expect better service if you choose to pay for it rather than do it yourself. Moreover, the metrics used for service levels will typically be different in a cloud environment. This discrepancy will vary from enterprise to enterprise depending on the expectations of each. In most cases, it doesn't matter how well the service provider delivers against its own targets because the customer may find it difficult to relate those targets to their internal KPIs. Service providers need to communicate in relevant business terms to manage expectations and, therefore, build trust.

managing end-user perception

An end user will wait weeks or months before reporting a performance-related problem (or may never report it), *unless* it means that he cannot access a business-critical application that he needs to do his job. The performance brown-outs may vary depending on *when* he uses the application or *which* device he uses to access it. During the downtime, he won't be as productive as he would otherwise be and will probably share his dissatisfaction with his colleagues. This negative build up around a particular application will most likely escalate if it is delivered by an external service provider, who very quickly becomes pinpointed as the cause of poor performance. *The end-user experience must be consistently delivered to an accepted performance benchmark for the service to be deemed reliable, and any degradation must be automatically detected as end users may not report it*.

managing performance as a component of availability

In traditional architectures, the internal support organization is accountable for the end users' experience of the IT service being provided. When that service is outsourced, most of the operational risks are transferred to the service provider – allowing the enterprise to shift any issues affecting its daily activities to the service provider and to potentially recover lost revenues (perhaps due to a reduction in business function throughput) through penalties. As the enterprise's main focus is on the availability, confidentiality and integrity of the service, it is important to consider that *in the eyes of the end user, a service is not considered available if performance issues are repeatedly experienced*. These end-user-experience considerations are not always scoped into service level agreements (SLAs), which can lead the customer to become despondent and unsatisfied with the service, even if it is delivered against agreed SLAs!

why should the service provider care?

In a strongly contended cloud computing market, the service provider will need to find a point of differentiation. As the market matures and cloud computing offers become more mainstream, it is expected that the most successful of the providers will be those who have built a partnership with their customers – and don't just sell CPU minutes, storage and bandwidth. The challenge is to develop a *partnership* based on trust.

traditional SLAs protect the service provider and do not build trust

Trust is based initially on transparency and is earned over time through shared experiences of reaching mutual goals. It requires that the service provider speaks the same language as the enterprise customer and understands the business requirements of the service – not just the IT requirements. Traditional SLAs based on technical KPIs for round-trip time and CPU minutes do not map directly to business requirements. The service provider needs to measure the performance of the cloud service in terms that directly map to the business requirements of the service – only then will the customer start to trust that the service provider is managing his best interests and not simply protecting itself with technical SLAs. For example: the staffing policy of a contact center is directly related to the number of calls that can be handled (i.e., the throughput), which in turn is based on the average length of time of a call, which in turn is dependent on the responsiveness, reliability and availability of the applications. Service providers need to design KPIs that map to business performance metrics and offer these as SLAs. *Fundamental to this are the end-user-experience KPIs, the measurement of which only the service provider can support, as it requires transparency of the end-to-end system through monitoring of the network traffic*.

guilty until proven innocent

The cloud computing service provider will – in the eyes of the subscriber – be considered "guilty until proven innocent" for any issues that affect the end-user experience. The service provider has implicit responsibility to isolate the cause of an issue to either absolve its responsibility or to fix the problem. If appropriate action is not taken (supported by tangible proof), the service provider risks gaining a poor reputation and poor customer satisfaction for issues for which it is not responsible. *Conversely if the provider can quickly isolate the issue and help the subscriber fix it – it stands to gain trust and build a partnership*.

change must be managed carefully or risk poor user acceptance and low adoption

The uneasiness of an enterprise to move to the cloud will be offset by rigorous User Acceptance Testing (UAT) for which tangible performance criteria must be designed around quality of experience (i.e., as seen from the eyes of the end user), not just the quality of the service. An inexperienced enterprise may choose to use stopwatches or page-load timers as criteria, but this data is typically inconclusive as it does not consider operational and business cycles or other external factors that are critical when operating in a shared environment. In this context, it is crucial that, at the onset of a migration to the cloud, the service provider insists on testing end-to-end performance based on end-user-experience metrics (to get the complete picture and establish a baseline), as any issue not identified at this stage will automatically be assigned to them.

impact to performance of dynamic workload management and volume elasticity

As the service is scaled up, virtual machines (VMs) are moved across physical hosts, and demand is increased in the cloud to support additional volumes of users, it is critical that end-user experience remains consistent with the original benchmarks. Without the support of the service provider to provide the transparency achieved with end-to-end monitoring, the adoption and growth of the cloud service in the enterprise is at risk. Capacity must be managed in coordination with performance management, and only the service provider can instrument these requirements.

challenges affecting the service provider

Organizations are becoming more defiant of their service providers. Traditional SLAs based on technical metrics (i.e., in the network world of RTD and packet loss) mean little to them and, in fact, are interpreted as a means to protect the service provider rather than ensure that adequate service levels are provided to their business. Mature enterprises are implementing service level management (SLM) toolsets in-house to manage their vendors. As an enterprise matures in this way, it becomes less reliant on service providers and, with the increased flexibility, can afford to make procurement changes more frequently. As a majority stakeholder in the end-to-end service delivery chain, the telcos can differentiate themselves from public hosting services by offering end-to-end SLAs. The challenge is to define these SLAs with relevance and therefore be of interest to business subscribers.

the enterprise must trust the service provider's instrumentation

The challenge for the service provider is to build strategic partnerships and reflect a position of mutual benefit when agreeing service levels with customers. Service level management is a function that should be provided by the service provider but cannot be conducted successfully without amicable participation with the subscriber. The alternative, whereby the subscriber takes service level management in-house and uses it to generate proof to prosecute the service provider, only leads to disputes about whose tools are measuring correctly – which slows issue resolution and leads to low customer satisfaction. The risk for service providers who have based their business models on a high-risk approach (i.e., have not invested in a resilient infrastructure and reactively purchase capacity), is that they stand to be exposed by the level of transparency delivered by service level management. The service providers who have maintained a low-risk approach (i.e., invested in developing a resilient and redundant infrastructure with back-to-back SLAs throughout the support chain albeit at a higher price point) stand to benefit from the trust they will build with their end customers.

enterprises have application performance management tools and WAN optimization already

The service provider must deliver (or support in an infrastructure-as-a-service model) the same level of application performance achieved by the customer's in-house application delivery architecture.

telcos must differentiate by offering end-to-end SLAs

The analysts have not yet made up their minds about the benefits of subscribing to a cloud computing service from a telco. This clearly demonstrates how immature the market is but also the potential for the service provider. Telcos see cloud computing as a natural progression and extension of their network-based services. This agrees with the view that cloud computing is not merely a platform but an enabler embedded within the enterprise architecture. Only the telcos with global reach and mature monitoring tools and processes can provide end-to-end service models.

SLAs must be relevant to the business

As stated above, the challenge for the service provider is to define SLAs that are appropriate to the level of business criticality of the cloud computing service. The risk profile (i.e., business exposure when application is not available) will determine what type of KPIs are required to manage the service. The service provider needs to offer options based on these KPIs, whereby SLA metrics based on end-user experience will be attached to premium-based products and not necessarily to entry-level cloud computing services. The challenge therefore is to define an attractive price point for SLAs based on end-user experience, which must also take into consideration that these SLAs will help the service provider foster a trusted relationship. A trusted relationship like customer satisfaction has financial value and should therefore be considered an investment opportunity.

recommended response

Service providers must manage the new risks to customer satisfaction by addressing the underlying technical, organizational and psychological issues of the cloud computing service model. Focus should extend past the customer's support organization requirements to include consideration of the business impact, i.e., how the end user "experiences" the service and the related impact to business productivity, brand, margin and revenues.

service providers should embed application performance management into the cloud computing service lifecycle to support SLAs based on end-user experience metrics

Irrespective of which cloud computing service model (i.e., infrastructure as a service, platform as a service or software as a service), the success of the service in the eyes of the subscriber is dependent on how the end user "experiences" the service. These experiences amount to how responsive, reliable and predictable the application is and are affected by how quickly any deviations from the norm are resolved.

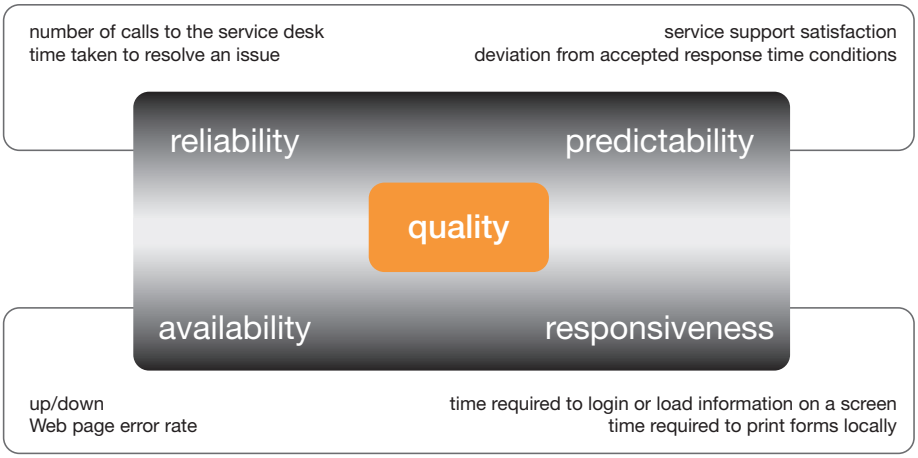


figure 2: end-user experience KPIs and metrics

An effective way to ensure that the service is orientated to support these success factors is to design the SLAs around end-user experience measurements. These experience-based targets must be defined by the business and map to targets for employee productivity, sales conversion ratios and customer satisfaction. This will mutually benefit both parties within a short period of time as each will have a view of how the technology maps to the business and will provide the transparency needed to build a valuable trusted partnership.

The disciplines of application performance management (APM) are defined as processes and the use of related IT tools to detect, diagnose, remedy and report on service levels to ensure that the cloud computing service meets or exceeds the end-user and business expectations. Application performance management

processes must be embedded in all of the lifecycle phases to ensure that the technology platform has the correct capacity and is built with the appropriate resilience, the components support the required performance specifications, and all components are configured and coded optimally. A proactive approach is mandatory in a dynamic and volatile cloud computing ecosystem when these capacity, performance and configuration requirements are constantly changing.

lifecycle phase	risk management process	application performance management processes
assess	identification and classification	assess existing state user experience benchmark and simulate performance modeling to support cloud migration
design	mitigation planning	define business level KPIs and map to SLAs define APM solution requirements
implement	control	implement APM solution
manage	monitor controls	perform monitoring as a component of service management
optimize	re-identification, classification and mitigation planning	redefine the requirements and/or recalibrate the performance model of the APM solution as part of the service level management process

table 1: risk and APM process management

Only through embedding application performance management processes and tools into the cloud computing services model can end-user metrics be measured and the necessary level of transparency in the system be achieved.

a consultative approach to manage risks and jointly define relevant service level agreements

The level of business exposure caused by an application performing outside of acceptable tolerances needs to be managed by identifying the risk profile and establishing controls to protect the business from losses.

A risk management approach will identify the potential impact and likelihood of a service outage or brownout (i.e., a performance issue that renders the service unavailable or degrades the business throughput).

risk = impact x likelihood

The impact of any service outage is linked to (1) the business criticality of the cloud service, and (2) the length of time it takes to resolve and recover from the issue. For example, the financial losses for every minute, hour and day of downtime.

impact = business criticality of the issue x time to resolve the issue

The likelihood of a service brownout (i.e., the probability or the expected number of times one will occur) is linked to a number of factors:

- complexity of the ecosystem (i.e., on-net vs. off-net vs. Internet/application architecture/end-user devices/geographic topology, third-party partner dependencies in the delivery chain)
- scale of the ecosystem (i.e., the number and geographic topology of users, the number of servers and tiers, the number of data center locations or peering points)
- reliability of the application pre-cloud transformation (i.e., the number and nature of service desk calls related to availability and performance and end-user satisfaction ratings)
- sensitivity of the application to external factors (i.e., packet transmission loss, variable latency and bandwidth within different application tiers and final delivery to the end point)

For example, the typical risk profile will be higher for a complex front-office application than for a simple back-office or IT function. Although, if due to the known complexity of the front-office application, the organization has a large support team with a track record of fast issue resolution, then the impact, and therefore the risk profile, may be lower than the back-office IT function.

Based on the risk profile, the service provider and subscriber can jointly agree on the appropriate level of control required to assure the service level requirements. *In the context of application performance management, these controls are tools and processes integrated within each phase of the service lifecycle to proactively manage the vulnerabilities of the relevant service and support KPIs.*

The response to a service's risk profile identifies which KPIs and which application performance management processes are required to support SLAs appropriate to the business exposure.

business criticality of application	complexity of application	type of cloud computing service	type of business	response
back office	low	private	SMB	accept
back office	low	premium	SMB/MNC	reduce
front office	high	private/ premium	MNC	share

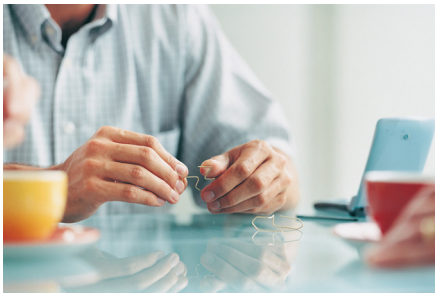
table 2: service level requirements response matrix[1]

1. accept the risks and establish the service based on technical KPIs only
2. reduce the risks by establishing application performance SLOs or SLAs based on network and platform performance metrics
3. share the risks by establishing SLOs and/ or SLAs based on end-user experience and linking the end-to-end service assurance with business metrics to provide service level management

In taking a consultative approach based on application performance management processes and tools, the service provider will facilitate the growth of a partnership in which both parties work jointly to identify and quantify the risks specific to the business service quality requirements, thus ensuring appropriate service levels are established in the best interests of both parties.

[1] response matrix table definitions
SMB: assumed low geographic dispersion of end users
MNC: assumed high geographic dispersion and high concentration of end users private: dedicated platform (no contention for resources with other organizations)
premium: shared platform (higher contention for resources and higher concern about resources being shared fairly among subscribers)
front office: business-critical applications (SAP, Oracle forms) back office: IT service functions (back-up, storage, email)

rationale



Remembering that *the cloud is an enabler and not the end goal* is key to realizing that the success of the cloud computing service is measured in business terms and not technology metrics. This means that a *common denominator* capturing the dependencies between the technology and the business must be found. This common denominator will then form the basis of service levels used to measure the effectiveness of the underlying technology of the cloud computing service.

The relationship between the business and the underlying technology can be expressed in terms of the quality of end-user experience … the common denominator.

CPU-minutes and RTD metrics are used for billing, while the success, and therefore adoption, of the service is based on the quality of experience

The subscriber is directly concerned with service availability, performance, confidentiality and integrity. Availability and performance are intrinsically linked as a performance brownout renders the service unavailable. The subscriber is more interested, therefore, in service levels that reflect the speed with which the service provider guarantees to resolve availability issues. It can be argued that the lower level metrics, which describe platform utilization and capacity, are only of indirect interest and are used to size the underlying service platform. Remember that cloud computing is elastic; if utilization is high, then additional capacity can be added to meet the needs of the business. Put simply, the subscriber will use metrics of CPU minutes or network bandwidth reservations to *size* the service (and therefore expect these to form part of the billing constructs). But the subscriber will measure the *success* of the service in terms of operational metrics based on their experience of the end-to-end service. By measuring metrics related to end-user experience, the service provider will be poised to detect and resolve issues that really matter to the subscriber – quickly.

subscribers need to know they are getting their fair share of the shared resources

Trust is earned by demonstrating one's understanding of the impact of the cloud computing platform on the customer's business – but, that's not all. The subscriber of a shared cloud platform is also concerned about the impact other subscribers could have on his system. In the same way that public Internet users may be concerned about speed, performance and security issues like confidentiality and integrity, a cloud customer will want to make sure that the resources to which he has subscribed will actually be his and not offered to or affected by someone else.

only the service provider can provide end-to-end monitoring

The level of visibility required for end-to-end performance management can only be provided by the service provider. For sure, a customer can install robots on the edge of the network to create synthetic transactions to measure the availability of certain application processes and their relative performance, but the instrumentation required to measure the real user experience end to end can only be achieved by the service provider at strategic locations within the network. The subscriber has no option but to use the service provider's performance management solutions. It is therefore in the service provider's interest to integrate application performance management processes and tools into cloud computing services.

the way forward

While organizations recognize that cloud computing services reduce the operational overhead of managing IT infrastructure, there remain two key barriers to adoption: can the organization trust the service provider to deliver a secure service that meets service level requirements that are relevant to the business?

security

Much of the risks surrounding data confidentiality and integrity can be mitigated by extending already well-established processes and tools into the cloud computing services and achieving certification to compliance benchmarks. The third component, “availability,” must be expressed in terms of up/down and performance metrics for which end-to-end visibility of the application chain is required.

service levels based on relevant business metrics

The challenge in the application delivery chain has expanded from delivering data from site A to site B to delivering an *application* to an end user. The service provider’s accountability has changed from delivering services to an organization’s *IT department* to delivering services directly to an organization’s *end user*. The service provider must therefore evolve to offer services based on relevant end-user experience metrics.

Application performance management is a fundamental component of cloud computing services, enabling end-to-end transparency to support both security and service-level requirements. Perhaps indirect and more subtle are the benefits that application performance management brings to the service provider-subscriber relationship of closer intimacy and trust. Only when all parties “talk the same language” (i.e., when the business impact is understood) and are motivated by the same targets will the shared responsibility for delivering cloud services be mutually successful.

about Orange Business Services

Orange Business Services, the Orange entity for business, is both a telecommunications operator and IT services company dedicated to businesses in France and around the world. Our 20,000 employees support companies, local government bodies and public sector organizations in every aspect of their digital transformation. This means we're at hand to orchestrate, operate and optimize: mobile and collaborative workspaces; IT and cloud infrastructures; connectivity (fixed and mobile networks, private and hybrid systems); applications for Internet of Things, 360° customer experience and big data analytics – as well as cybersecurity, thanks to our expertise in the protection of information systems and critical infrastructures. More than 2 million businesses in France and 3,000 multinationals place their trust in us. See why at: orange-business.com or follow us on Twitter @orangebusiness

Business Services

orange™