EXECUTIVE SUMMARY

Enterprises are looking today at consistent end-to-end application performance guarantees. Central to many business applications, the WAN is a crucial component to achieve that goal. Service providers must therefore offer some guarantees in terms of the performance of applications running over the WAN.

Offering such guarantees is the role of Service Level Agreements (SLAs) that are defined between the service providers and the enterprises. But typical network services SLAs are not directly linked to the performance of business critical applications.

This paper takes a look at existing network SLAs and explains why they are not providing a complete answer to modern enterprises performance challenges. It defines the key requirements that need to be considered when putting in place application-centric SLAs. Finally, it outlines the key concepts behind Ipanema’s application-centric SLA framework.
INTRODUCTION

Enterprises are looking today at consistent end-to-end application performance guarantees. Central to many business applications, the WAN is a crucial component to achieve that goal. Service providers must therefore offer some guarantees in terms of the performance of applications running over the WAN.

Offering such guarantees is the role of Service Level Agreements (SLAs) that are defined between the service providers and the enterprises. But typical network services SLAs are not directly linked to the performance of business critical applications. For example, a network service may comply with SLA guarantees even though the enterprise might suffer from an important SAP module that performs poorly—because of network issues.

Virtual Private Network (VPN) services are evolving to become more application aware. Peter Hall, Enterprise Services Research Director of market research firm Ovum said, “a few service providers are making the first steps towards applications-centric SLAs, sometimes referred to as the Holy Grail of networking.” Moving from a network-centric model to an application-centric model is a great opportunity for service providers because it allows them to deliver higher-value services that support application requirements.

This paper takes a look at existing network SLAs and explains why they are not providing a complete answer to modern enterprises performance challenges. It defines the key requirements that need to be considered when putting in place application-centric SLAs. Finally, it outlines the key concepts behind Ipanema’s application-centric SLA framework.

SLA BASICS

Service providers face increasing pressure to offer high-value services that can be provided reliably and consistently. As enterprise customers continue to introduce applications over Multiprotocol Label Switching (MPLS) WANs, they become increasingly concerned about whether application performance can be maintained and optimized, particularly for mission-critical applications.

There are multiple types of network performance SLAs currently available:
- Reactivity SLAs define how the service provider is able to respond to an issue with the service.
- Availability SLAs define the ability of the service provider to continuously deliver the service.
Performance SLAs define the performance levels with which the service provider is able to deliver the service. This is the type of SLAs that are discussed in this paper.

For each SLA, key ingredients are required:

- **One indicator**: Service providers need to define the metric for measuring the SLA. For the sake of SLA simplicity, the number of indicator should be kept as low as possible, and the use of a single indicator is the most powerful solution.

- **Thresholds on the indicators**: Thresholds should be defined that represent the objectives of the SLA and define a range in which the service provider is meeting SLA commitments.

- **Rules for validity of the indicators**: They define the conditions on which the service provider is able to commit to the SLA.

Of course, a technology that actually delivers the required SLA is also a fundamental ingredient.

THE LIMITATIONS OF TODAY’S NETWORK SLAs

Traditional network performance SLAs that are implemented by a large number of IP VPN service providers are constrained by several limitations. A key source of these limitations comes from the nature of the indicators they are built upon and also from the way those indicators are measured.

The indicators in use for performance SLAs are typically link delay or loss information that are measured from enterprise router to enterprise router by pinging a remote router or using Cisco Service Assurance Agent (SAA) probing. The problem with pinging and probing is threefold:

1. **The resulting SLAs are by design not representative of the application performance but are merely representative of the link itself.**

2. **The SLAs are not representative of the performance from LAN to LAN.** The ping or SAA measurements that in most implementations take place between Customer Premises Equipment (CPE) routers or shadow routers are not fully taking into account congestion issues within the enterprise.

3. **The SLAs are only an estimate of the performance, since ping tools and SAA are based on simulated traffic that takes place periodically.**
SLAs produced by pinging and probing are in no way connected to what is truly experienced by end-users of WAN applications.

1. Network SLAs Are Not Representative Of Application Performance

The result is that network SLAs by design are not representative of application performance but are only representative of link performance. A successful Application SLA must use metrics that go beyond the measurement of the link performance and connect to the performance delivered to the end-users of applications.

Working at the level of individual traffic flows and individual applications makes it possible to gather a number of key performance metrics. For TCP applications, the Round Trip Time (RTT) is a good indication of the performance of the flow. The delivered throughput is also highly connected to the end-user experience. A deeper observation of the TCP protocol exchanges can deliver an ever better indication of the performance delivered to the user. For example, the transaction time can be computed by observing client requests and server responses at the TCP level.

For UDP-based applications, there is no equivalent to RTT or transaction time so metrics such as for one-way delay jitter and losses must be used. These metrics are much more difficult to collect since they require advanced measurement capabilities. To compute one-way delays or losses, two probes must be involved; one near the client and one near the server. For TCP metrics, one probe near the server is enough and the technology involved is also simpler.

2. Network SLAs Are Not Representative Of LAN to LAN Performance

For the Application SLA to be meaningful to enterprise customers, it must clearly express the performance delivered by the service to the end users. It must also accurately represent the portion of the application delivery chain that the service provider controls. In most cases, a service provider’s control stops at the enterprise; the service provider usually has no responsibility for the behavior of the LAN, the servers or the clients. Metrics computed between a router on one LAN to a router on the remote LAN define precisely the portion of the application delivery chain that the service provider controls.

Not all application flow metrics have the ability to measure exactly what is going on from LAN to LAN. For example, the RTT can be high if there are performance issues on the network—or on the enterprise infrastructure. As a result, Application SLAs with an indicator based on the RTT are dangerous for service providers because the measured performance can be poor without the service being at fault. For example, the enterprise could be experiencing a bottleneck at a server or a router attached to the LAN, and the service provider could unjustly bear the blame.
One-way application flow metrics computed from LAN to LAN—such as metrics for throughput, delay, loss and jitter—represent a solid choice for Application SLAs. They apply to all applications—both TCP-based and UDP-based—and precisely cover the aspect of the infrastructure that is within the control of the service provider.

3. Network SLAs Are Only Estimates Of Performance

The Application SLA will have a strong value to the enterprise if the indicator on which it is based covers what is being—or has been—truly experienced by end-users. A periodic measurement of simulated traffic is not representative of the actual experiences of end-users.

The primary problem is that even the service provider will not be able to know what quality level has really been delivered. If performance degradation occurs from an end-user’s standpoint, the enterprise is likely to immediately point to the service as the cause of the problem without any one being able to verify the root cause of the issue. This inevitably leads to the enterprise unjustly blaming the service for an enterprise performance problem that the service provider has no possible control over.

Monitoring each single flow accurately is key to bringing the highest value to an Application SLA. This allows service providers to irrefutably prove the quality being delivered and therefore establish greater trust with enterprise customers. Service providers deliver increased value to the enterprise, since they can help customers accurately identify when performance problems are actually within the enterprise so organizations can better support the consistent delivery of application performance to end-users.
INGREDIENTS FOR AN APPLICATION SLA

Any accurate, performance Application SLA requires:

1. **One indicator**
2. **Indicator thresholds (the objectives)**
3. **Rules for validity of the indicators**

These three ingredients must of course be complemented with a technology that actually delivers the SLA. This section describes Ipanema’s solution for efficient enterprise Application SLA.

1. **Choosing An Indicator**

The previous section outlined the reasons why network SLAs are not enough and offered hints on how to overcome their limitations. An indicator suitable for Application SLA should be representative of the:

- Application
- LAN to LAN performance
- Real quality levels delivered by the service provider

The only basic metrics that comply with these requirements are one-way application flow metrics computed from LAN to LAN for throughput, delay, loss and jitter. But those metrics might be too numerous and too technical to handle in the Application SLA. It is important for an SLA of any kind to manipulate a small amount of meaningful indicators—and ideally only a single indicator.

For VoIP applications, the International Telecommunication Union (ITU) is proposing a model—the E-model—that allows service providers to transform low-level, one-way metrics into a single high-level score: the Mean Opinion Score (MOS). The MOS was created to estimate the delivered quality of voice services using subjective testing. Users conduct test calls and rank the sound quality on a scale of 1 to 5. Later, with the advent of VoIP, the ITU proposed the E-model that takes “flow impairments” such as delay, loss and jitter and Coding/Decoding (CODEC) technology to compute an intermediate factor—the Rating (R) factor that directly translates into a MOS value.

For other applications such as client-server applications or even video applications there are no standard indicators. The Application Performance Index (APDEX) was developed by the
APDEX Alliance in an attempt to standardize a MOS-equivalent for client-server applications. It proposes to measure the response time obtained by end-users to perform typical tasks through their applications. The APDEX framework is interesting but is not usable with Applications SLAs.

APDEX makes it extremely difficult to isolate the portion of the application delivery chain that the service provider actually controls. Computing APDEX by only looking at the traffic flow is awkward, since it requires an understanding of the applications semantics, including their different versions and flavors as well as an understanding of the standard behavior of users in their typical business processes and application tasks.

The need for an equivalent of MOS and the limitations of APDEX are the reason why Ipanema developed the Application Quality Score™ (AQS) as an indicator of choice for Application SLAs covering applications other than VoIP. The AQS is built on the same principles as the E-model. It transforms basic application flow metrics—such as measurements for throughput, delay, loss and jitter—into a high-level score using a proprietary model. This model is based on the comparison of the multiple application flow characteristics with thresholds stored in a predefined library.

Figure 2: The LAN-to-LAN « Application Quality Score™ » (LL-AQS)

A first step computes an intermediate value: the Red Amber Green (RAG) indicator that simply defines whether or not a given flow complies with the minimum amount of resources required for a good experience for end-users. Green means the flow has this minimal amount of resources. Amber means that it does not have all the optimum resources but enough for an acceptable quality of experience. Red means that the flow does not have the required resources at all.

The RAG indicator is then collated across and between sessions to obtain a numerical value between 0 and 10 using a weighting associated with the percentage of data delivered to end-users with Red/Amber/Green RAG indicators. Both the MOS and AQS make ideal Application SLA indicators when used with precise, exhaustive measurement of the real traffic from LAN to LAN. In such a configuration, the MOS and AQS are referred as L2L-MOS and L2L-AQS.
2. Defining Thresholds

A straightforward way of defining Application SLA thresholds is to build an analogy between network availability and application availability. An organization can consider that an application is available to end-users as long as the network service is able to deliver it with the appropriate resources during a defined period of time. The commitment to the SLA can be delivered over the whole network or by site. The commitment by site has more value since an SLA commitment across the whole network can hide minor site-level SLA breaches.

It is also important to consider that since the network is shared among many applications, the notion of Application SLA should be positioned to support critical applications only and used in conjunction with some mechanisms allowing enterprise customers to intelligently allocate network resources to users of critical applications in priority. For the sake of simplicity, applications of the same criticality to the business can therefore be regrouped together in the same SLA. For example, an Application SLA would be defined for all the “Top” criticality applications, and another for the “High” criticality applications.

3. Rules For Validity Of The Indicators

The rules for validity of the indicators define the circumstances under which the service provider will be able to commit the Application SLA. Establishing rules for validity of the indicators is essential, since network resources such as bandwidth are finite. In the vast majority of cases today, a site’s network connectivity is defined by its fixed WAN access speed. The most basic rule for validity of an application SLA based on performance is simply that the SLA cannot be committed during periods when users of the guaranteed applications require more WAN resources than the enterprise has provisioned for the site.

Multiple refinements exist to define those periods when users of the guaranteed applications require more resources than what is available for the site. The first idea is to define a maximum number of simultaneous users for each application—or group of applications—covered by Application SLAs. If the amount of bandwidth is exceeded the service provider is not able to deliver a guarantee of acceptable quality levels delivered to end-users. Ipanema is proposing to use the notion of “over-activity”. Over-activity defines with extreme precision cases when users of critical applications are asking more of the network than it can realistically deliver. Over-activity combines the analysis of what:

- Users are asking for
- Their needs are
- The network is able to deliver

Over-activity can be computed in real-time and can feed both the SLA reporting and troubleshooting processes. For example, in an over-activity period the average AQS delivered in monthly SLA reports will not take into account the values of the period. On the other hand, SLA troubleshooting reports will highlight the over-activity period in order to recommend an
appropriate action to handle this over-activity, which is typically a recommendation to upgrade the link.

4. Application SLA Example

Below is an example of an Application SLA defined with Ipanema’s approach. Applications are grouped into four different criticalities Top, High, Medium and Low. Top-criticality and High-criticality applications receive an Application SLA using the AQS, and Medium-criticality and Low-criticality applications are not associated with any SLA. In this example, VoIP flows receive dedicated SLAs using the MOS.

In this example, the customer receives a monthly Application SLA report that provides meaningful information that allows the enterprise to accurately analyze the WAN’s impact on the performance of important enterprise applications.

**Figure 3: Monthly review of the SLA**
Ipanema Technologies offers a unique Business Network Optimization solution that automatically manages and maximizes network application performance according to how critical applications are to the enterprise. The Ipanema Business Network Optimization solution bridges the gap between the enterprise's business priorities and the Enterprise network infrastructure.

Enterprises want to get relevant SLAs on the key section of the application delivery chain that the network represents, and Ipanema is providing a solution that allows service providers to define and measure SLAs on the network that:

- Connect to the business of the enterprises through applications
- Clearly and precisely covers the different sections of the application delivery chain
- Are representative of the actual quality levels delivered to end-users

Figure 4: The Ipanema System is a traffic management solution based on distributed devices and a central software suite

Ipanema provides a turnkey Application SLA definition, measurement and enforcement framework that includes a range of high-level indicators suitable for Application SLAs, the
ability to automatically check for Application SLA validity, a set of dedicated SLA tools and reports, and the ability to actively enforce the Application SLAs over the network thanks to a global and dynamic approach to traffic management based on application performance objectives.

Business Network Optimization solutions from Ipanema allow service providers to overcome the limitations of traditional network SLAs and gain control over individual application sessions. They can measure individual flows instead of link performance, setup mechanisms that measure performance beyond the WAN and measure real user traffic without merely relying on probing.

Ipanema’s Business Network Optimization is a traffic management solution based on distributed devices and a central software suite. The synchronized Ipanema devices can be deployed at multiple enterprise locations to measure the network’s contribution to application performance for each UDP and TCP flows. Application SLAs can therefore be measured globally over the network. Metrics are aggregated in three dimensions—application, topology and time—and stored in a database, and results are displayed through interactive reports and drill-down monitoring tools.

Application SLAs can be automatically and globally enforced over the network. Organization can centrally define application performance objectives, and distributed devices perform continuous, real-time and global analysis of the traffic and application mix. Traffic management policies are computed automatically and updated each second.
Service providers can benefit from Ipanema’s Scalable, Application Service Level Architecture (SALSA) to scale application SLA measurement and enforcement to support their entire installed base of enterprise customers. Network operators can therefore offer managed services with clear and measurable SLAs.

Application SLA measurement and enforcement capabilities of Ipanema Business Network Optimization allow service providers to create a new breed of application-centric services. Application SLAs fully connects to the enterprise business needs and provide a unique opportunity for service providers to increase market share, revenue and profits.
Ipanema Technologies is a provider of advanced application traffic management solutions that maximize network application performance. Its solutions are utilized by organizations deploying VoIP, ERP applications and other business critical projects as well as those embarking on ITIL projects. Network integrators market the Ipanema system to enterprises, while telecom service providers and network managed service providers offer it as a service.

The Ipanema system is simple, automated and scalable and is unique in guaranteeing critical application performance regardless of network conditions. It enables this by providing full visibility of application flows over the network, global and dynamic optimization of network resources and transparent application acceleration. Ipanema’s solutions are deployed in more than 75 countries. For more information about Application SLA solutions from Ipanema, visit www.ipanematech.com.