

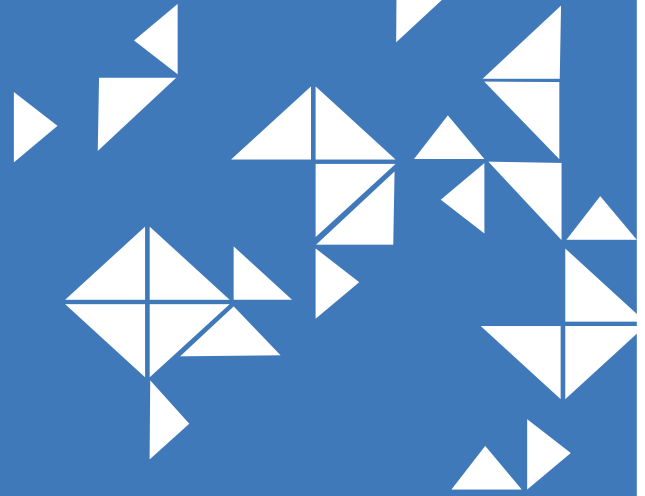


Enterprise White Paper

Application Performance Optimization: Challenges and Solutions

Louis Faraud, Simvar Lach

www.tatacommunications.com



Tata Communications
"Application Performance Optimization: Challenges and Solutions"

Table of Contents

Executive Summary	5
New Challenges for the WAN	6
Visibility	7
WAN Optimization	8
WAN Acceleration Solutions	9
Differentiated Service and Reporting	13
Technical Overview	14
Key Factors for Choosing the Optimization Solution	15
Tata Communications' Application Performance Optimization Solution	16
Case Study	18
Conclusion	19
Acronym Key	20
Contacts	21

Executive Summary

The Forrester Research Wave report of July 2007 titled "WAN Optimization Appliances, Q3 2007" defined the WAN optimization market as "red hot." Global companies are putting their networks under increasing stress, and data exchanges to and from centralized data centers have dramatically grown as legacy client/server protocols have transitioned to Internet Protocol. For distributed organizations, facilitating collaboration within their global workforce is a daily challenge that, at least partially, rests on their ability to share information. This increasing traffic overloads the enterprise network, making the WAN congested, slow, and error-prone. Thus, WAN optimization needs to be a priority.

Understanding the challenges faced by companies and the current optimization solutions is the key to successfully adapt the company network performances to their applications. This White Paper will outline:

- _ New performance challenges faced by companies within their WAN
- _ Existing optimization and acceleration techniques
- _ Vendors' solutions
- _ Tata Communications' offer: Application Performance Optimization

New Challenges for the WAN

Distributed enterprises' networks are bogged down by significant data traffic, leading to congestion caused by common patterns of network usage. Growing data exchange consumption is generally not being offset by a corresponding increase in bandwidth.

Performance requirements

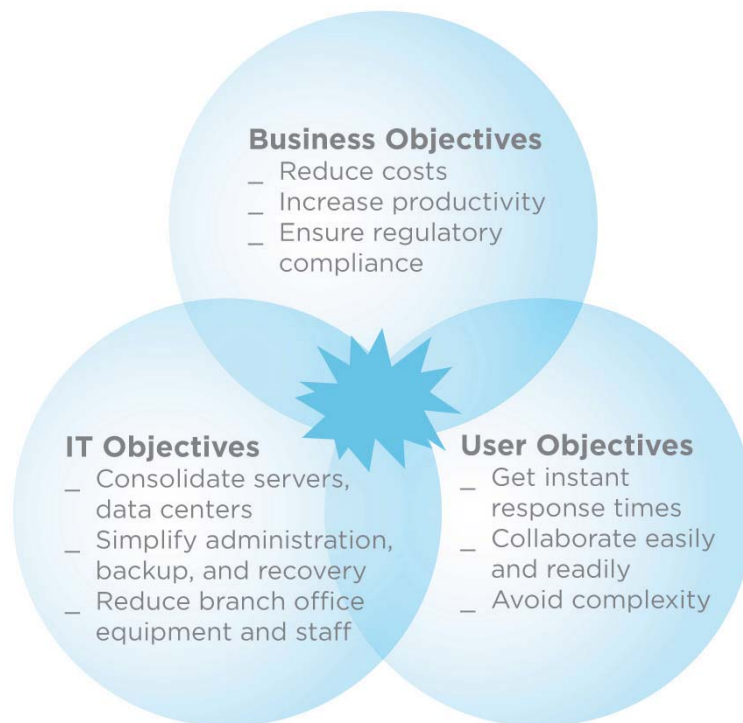
WAN performance requirements generally include:

- _ The need for data mutualization. The implementation of strategic cost-reduction solutions has led to an initiative to consolidate storage in large centralized data centers rather than host it locally.
- _ Increase of bandwidth usage. The "Webification" of applications is now glutting the network, since web-based solutions are more bandwidth-consuming than their client-server application predecessors.
- _ Application and end users response time. The use of applications that were efficient in a LAN environment are not optimized to be used in a distributed architecture, causing abnormally large volume of data traffic on the enterprise WAN.

Resulting issues

The aforementioned challenges overload the enterprise network and can result in:

- _ Creating more network dependencies between the distributed workforce and centralized data servers.
- _ Increasing the amount of data exchanged and the total number of exchanges between end users and data centers, since file storage is now consolidated (creating a Content Delivery Network). CDN can include data such as sensitive company information or more bandwidth-heavy applications like video (for e-training, for example).
- _ Increased congestion, as the WAN connecting remote offices to data centers have to support the Common Internet File System (CIFS) protocol, which enables client systems to request file and print services from server systems. CIFS is based on the Server Message Block (SMB) protocol, which was designed for LAN, not WAN use, which can result in a negative impact on remote file access performance. Other applications like email or collaborative software can cause similar problems.
- _ Adding latency that hampers the WAN performance. Due to the generalization of distributed users (remote offices, scattered end users) accessing distant data centers by crossing increasing numbers of network segments, application performance is hampered by the retransmission of data packets to the next segment (application performance can be no better than the slowest link in the network).
- _ Packet loss, as congestion occurs. Latency and congestion result in saturated links as the network devices start buffering packets instead of transmitting data. TCP acknowledgments are not sent, and the TCP window shuts to free some bandwidth. If the congestion is not stopped, eventually the buffers will overflow, resulting in packet loss and decreasing TCP throughput.



Visibility

In lieu of getting a faster connection, there are many different approaches to optimize WAN connectivity. The most critical step is to get the best visibility (performance and application behavior), so that the most high-impact network performance optimization measures can be chosen.

A thorough knowledge of the applications running on the network is important and could be achieved through an audit to:

- _ Gain application visibility on the network
- _ Measure applications and report on their performance
- _ Determine bandwidth utilization per application
- _ Measure throughput per application and of the WAN
- _ Diagnose problems and issues

WAN Optimization

There are several service features that can fulfill network optimization requirements. Historically, each of them has been an appropriate solution, chosen to improve a network's poor performance at a given time.

Control/Optimization	Acceleration
Traffic Marking and Classification	TCP Acceleration
Bandwidth Allocation	Packet Packing
Priority Handling	Compression (Disk and RAM based)
Policy and Shaping	Caching (Disk based)
Forward Error Correction	
QoS	

Optimization can be defined as intelligent management of traffic flow between the branch office devices

Network Performance and Quality of Service

Quality of Service agreements combine all the protocols and procedures needed to control data network performance end-to-end. Maximum values for performance elements like transit delay, jitter, or packet loss are chosen according to what is known about application requirements or user needs so that business objectives are more readily achieved.

Corporations that outsource their networks to service providers use Service Level Agreements (SLAs) to guarantee performance of the transport network and provider service delivery. The performance element used for response time SLA is usually measured with an IP echo command between customer network sites. According to delay variation due to packet queuing at the ISP's Customer Edge Router, transit delays are collected under predefined conditions: the Transport Network Provider cannot be responsible for bad performance if the tail line is overloaded by customers' traffic.

Class of Service

For several years, network solutions have included mechanisms to prioritize critical applications, allowing them to meet performance requirements even when links become overloaded. Several classes of service (real time, urgent, best effort,) have offered remote sites the ability to differentiate network traffic. Non-bursty, low-bandwidth applications can receive priority treatment when the access line is overloaded.

Specific Quality of Service agreements (jitter, transit delay, packet loss) have been implemented on a per class basis instead of on a per site basis. According to queuing theory, around half of the tail line bandwidth can now be reserved for committed classes of service. If applications are not compliant with Class of Service rules, routers will randomly delete packets (in excess), to force session control mechanisms to reduce the application throughput.

Class of Service processes have addressed the problem of limited bandwidth and led to a compromise that gives vital applications a minimum portion of the resource in scarce conditions. But there is still a need to optimize for high bandwidth consuming applications with LAN-like transit delays for the distributed enterprise.

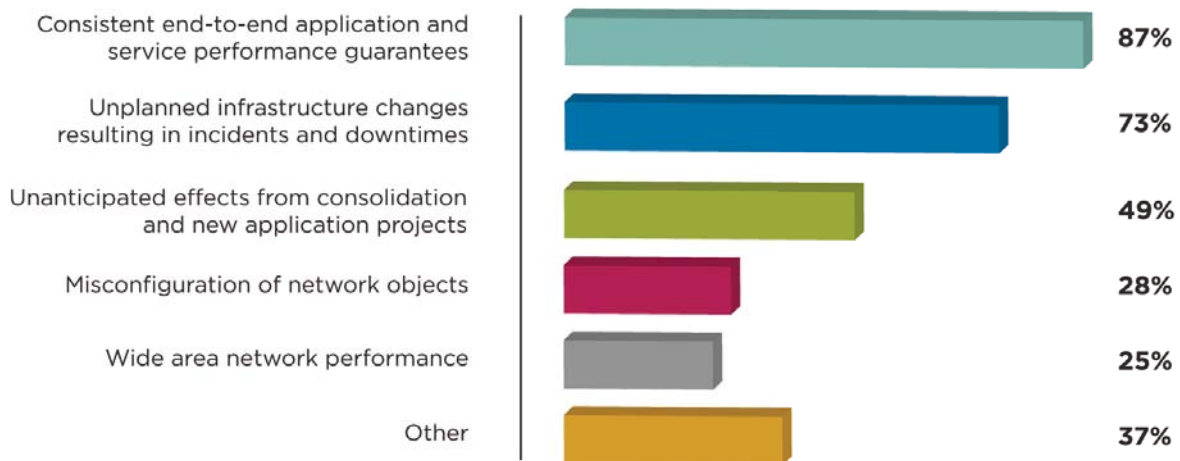
WAN Acceleration Solutions

Companies are now looking for solutions to go beyond Class of Service and enable bursty, high bandwidth applications like HTTP, FTP, CIFS or Samba, to ensure consistent end-to-end committed performance without using "unlimited bandwidth."

The three major concerns studying are typically:

- _ Data caching
- _ Data compression
- _ Protocol acceleration

"What are your top three issues in managing corporate IT infrastructure?"



Base: 67 enterprises IT infrastructure managers at \$1 billion-plus companies (multiple responses accepted)

Source: Forrester Research, Inc.

Data caching

The best way to ensure low transfer time when accessing a data file is to access data stored on a Local Area Network. By using File Transfer Program (FTP), customers have the ability to transfer data from one FTP server to another. This is called "setting up a proxy." If this proxy server is locally attached, it can greatly reduce transit delay. The "proxy get" and "proxy put" commands can be used to push data from a centrally hosted server once during the night, and make it accessible locally several times during the day. This process will minimize transit delay and reduce the amount of traffic within the client network.

The goal is not to re introduce all local resources, but rather to install the minimum appliance to set up an automated process to store a copy of recently accessed data on a local storage unit. If this process is adapted to a great number of network applications, it would be of great value in saving bandwidth.

Caching is done in specific hardware often called BOB (Branch Office Box) or WOC (Wan Office Controller), which is usually installed in series on the LAN side of a Customer Edge Router. The CISCO hardware platform, Wide area Application Engine (WAE) can be installed in parallel: the targeted traffic is diverted from the router to the hardware platform. Alternatively, the hardware platform can be integrated into a router's slot.

When a data file is first accessed, the Branch Office Box interprets protocol-specific dialog, the needed data is retrieved from the server, stored locally and then delivered to the end user. This process can be efficient only if the retrieved data is reused during the period the data is stored in the box. For the first user accessing the data, the turnaround time actually degrades because the cache process inserts an added delay to the end-to-end transaction. WAN Office Controller providers have improved the initial access to the file by implementing Data Compression and Protocol Acceleration.

Data compression

One of the first responses developed to slow transmission times was data compression. The first method deployed suppressed repeated characters to deliver an average compression rate of 30%, which increased significantly when users sent pages of blank characters. This sort of compression is done at the level 2 protocol between two routers directly connected.

Lempel-Ziv compression algorithm is a more efficient way to lower the amount of data. A dynamic dictionary running at each end of the data path will maintain a list of data strings and replace them in the transferred file by their corresponding references. This technique helps to reduce data by more than 50%, if the data has not already been compressed by an equivalent technique, and is run end-to-end between two BOBs.

Data Redundancy Elimination is the most effective compression algorithm, reducing the data by more than 80%, according to manufacturers. Traffic patterns across several sessions are detected and stored at the two ends of the data path in the compression boxes.

The process would not be complete if improvements were not implemented in IP communication protocols.

Protocol Acceleration

Protocol Acceleration takes place at the transport layer of the OSI model. TCP and UDP are the main protocols that provide end-to-end data transfer. TCP is connection-oriented and is able to adapt itself to the quality of the Internet layer. It provides flow control and error recovery. UDP (User Datagram Protocol) is connectionless, and doesn't provide flow control or error recovery. UDP is used by applications and provides their end-to-end flow control. FTP uses TCP at the transport layer, while TFTP (Trivial FTP) uses UDP. The usage of TCP means that the application may force its flow control: CIFS application is an example.

A TCP window is the amount of data a sender can send on a path before they get an acknowledgment back from the receiver. To test network performance, TCP sends increasingly larger windows until it sends one that's too big to be processed properly, and the acknowledgment isn't sent. The packet loss usually stems from network congestion. TCP resends the lost packet and decreases the sending window, adapting the flow control to the network capacity. In the case of a very slow network, TCP can retransmit a packet and reduce the sending window before it receives the acknowledgment packet. This is called the "silly window problem."

Protocol acceleration will increase the sending window at accelerated rates at the beginning of the session. The local WOC acknowledges the sent datagram and is responsible for delivering the datagram to the other end of the network. To avoid resending of lost packets, data is added to the datagram for Forward Error Correction (FEC) at the receiving WOC.

Data transfer between WOCs can be transported through a proprietary tunnel, such as IPSec. The packets are transparent to the CoS and QoS mechanisms of the transport network. When packet marking and transport sessions are preserved by the WAN Optimization Controller, the boxes are said to be transparent to the transport network.

Differentiated Service and Reporting

WAN optimization can't be studied without addressing differentiated service features and reporting, in order to ensure that Service Level Agreements are met and bandwidth management is adapted to traffic profiles.

Differentiated service

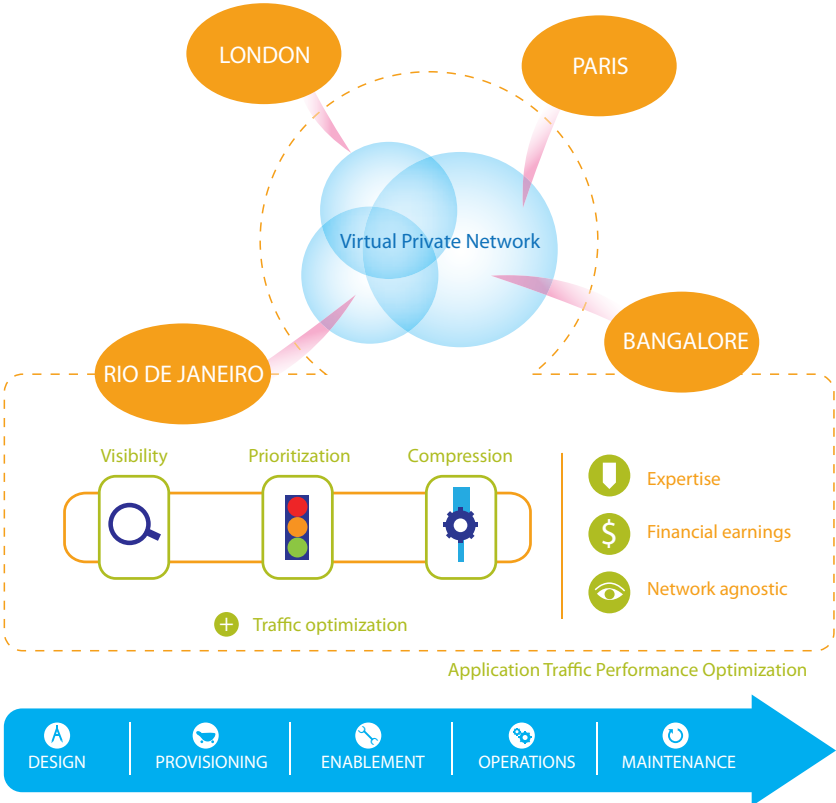
The traffic optimizer, as a Customer Edge Router, sees all the traffic going to or coming from the WAN. Differentiated services algorithms can be run in this box and relieve the router of that duty. Because the algorithms carefully analyze the IP transport protocols, they can increase the granularity of the prioritization process. They can identify an Internet or an Intranet HTTP session by a port number or with a set of rules. Policy-based routing can divert the best-effort traffic to a low-cost link.

When using tunnels on the transport network, the traffic optimizer hardware can manage the connection link congestion, prioritizing packets for critical application in the transport tunnel. On a hybrid VPN using different administrative transport domains, this can be a way to cope with heterogenous SLA.

These rules can be programmed and stored on a policy server, broadcast to the remote WOC and adapted by the local or end-to-end conditions with automated processes.

Reporting

A traffic optimizer can play the role of a probe, analyzing traffic flow and storing detailed traffic characteristics at the IP, transport or application protocol layer. This data can be consolidated into a matrix showing the fulfilled performance objectives for a particular site.



Technical Overview

Major vendors at a glance

The following table is a snapshot of leading vendors' current offerings (as of January 2009). This information is provided solely as a summary of features and is not a comparison or endorsement of any of the below-mentioned companies.

Solutions are constantly evolving and vendors are continuously integrating new features into their solution packages.

Manufacturer	CISCO	IPANEMA	JUNIPER	RIVERBED	BLUE COAT
Product name	WAAS	IP engine	WX / WXC	Steelhead Appliances / Interceptor	Packet Shaper
Data Caching	YES	Planned	YES	YES	iShaper Extension for Caching functionalities
TCP Protocol Acceleration / Data Compression	YES	YES	YES	YES	YES
Transport Tunnel	NO Target traffic diverted to WAAS	IPComp	IPComp	NO	Proprietary
DSCP Transparency	YES	Configurable transparency	Configurable transparency	YES	NO
Traffic Monitoring	YES	YES	YES	YES	YES

Key Factors for Choosing the Optimization Solution

When choosing network optimization solutions, look for a provider that offers the following benefits:

Visibility: a service should provide a complete and accurate view of the network traffic, in real time; and offer written reports stating how efficiently the network is operating.

User productivity: a primary goal of WAN Optimization is to adjust the network based on the needs and importance of different applications that a company is running. Prioritizing critical traffic enables a better response time for applications and increases ROI. This is especially relevant for remote sites or sites located in countries where the network cost is high as they access central applications (SAP, People Soft, Oracle), or when accessing corporate applications through the Internet. Optimization results from effectively managing different and evolving technologies: dynamic bandwidth allocation, TCP optimization, specific applicative optimization, compression, caching, or acceleration.

Remote sites integration in the company network: the company remote sites are enabled with central applications at sensible cost, thereby optimizing development and support resources while supporting performance.

Optimal sizing of bandwidth: a service provider should enable bandwidth adjustments based on what is needed at the time, and how quickly the network is growing.

Cost savings: the necessary bandwidth control prevents over-sizing the network and saves money.

Expertise: WAN optimization solutions should be based on tested and constantly updated technologies and should be supported by expert staff

Neutrality: Finding a service provider that emphasizes net neutrality and technology tailored to a company's specific needs, over bandwidth provisioning and allegiance to a specific manufacturer is critical.

Tata Communications' Application Performance Optimization Solution

Tata Communications' Application Performance Optimization (APO) service addresses a company's need for more efficient network performance for their applications.

The service provides:

- _ Identification of customer-specific application behavior on the network
- _ Increased application performance on the network: transactional latency, file transmission time
- _ Prioritizing of network resources to handle critical business applications: SAP, Oracle vs. electronic messaging or web surfing
- _ Visibility through monthly operating reports, and measurement of key-metrics such as volume, transit delay, jitter and packet loss per application
- _ Integration of remote sites into the global company network, adjusting bandwidth as needed for optimal results
- _ A cost-effective solution that is managed by a team of experts who are committed to delivering the best optimization service

APO initially provides an accurate view on application flow. It then optimizes the network by adapting it to the specific needs of the supported applications, while achieving quality control.

Tata Communications provides a cost-effective APO solution complete with architecture design, installation of the appropriate devices, traffic analysis, implementation of the optimization engineering technologies and end-to-end management of the solution.

Product features

APO covers the services as described below.

Solution definition

- _ Understanding of the customer's environment and needs
- _ Snapshot of the existing network environment, analysis of the impact of applicative evolutions to the network, study of the different solutions taking into account the investment plan (expected improvements, financial aspects)
- _ Return On Investment (ROI) calculation
- _ Design of the solution
- _ Nomination of the project and customer care teams
- _ Pilot test on selected sites to validate and refine technical choices (platform, configuration)

Enablement

- _ Purchasing, installation and configuration of the hardware and software platforms
- _ Analysis of the applicative traffic and enabling of optimization technologies
- _ Definition of Service Level Agreement (SLA) depending on pattern of applications

Operation Management

- _ Hardware and software platform maintenance
- _ End-to-end monitoring service for the concerned devices
- _ Availability of the customer helpdesk 24/7 and the fault management
- _ Performance management and reporting
- _ Analysis of reports and advice regarding network evolution

Case study

A customer runs a worldwide 30-site MPLS network, with sites located in Singapore, Japan (Tokyo), China (Shanghai, Hong Kong), Thailand (Bangkok), Malaysia (Kuala Lumpur), Taiwan (Taipei), Korea (Seoul), Australia (Sydney), France (Paris, 2 sites), Belgium (Brussels), Russia (Moscow), Germany (Hamburg), Spain (Madrid, 2 sites), UK (London), Italy (Milan), Switzerland (Geneva), US (8 sites), Canada (Montreal), Panama, and Mexico (Mexico City).

Applications running over the network included Lotus Notes (mail), HTTP, FTP, CIFS, industry applications (i.e. for manufacturing, ordering, billing) and Intranet. The installation of Tata Communications' Application Performance Optimization service led to an average bandwidth savings of 66%; 2/3 of sites saved between 68% and 77%, and the site with the lowest performance reached a minimum of 55%.

Those results were equivalent to upgrading the currently available network bandwidth 300%, in addition to full visibility for network usage per application.

The customer is now using this released capacity to deploy a Videoconferencing and VoIP project.

The return on investment was roughly 13 months.

Conclusion

WAN optimization doesn't just involve saving money on bandwidth or compressing protocols; information technology has changed, and distributed data storage has evolved to concentrated large computing centers. Web technologies have become a universal user interface, and new applications require increasingly more bandwidth. Single users want to be able to perform large data transfers at any time during the workday.

Merely knowing what is running on their VPN is no longer an option for organizations, as bandwidth is an expensive resource. A WAN optimization strategy should be made part of an overall data management plan - knowing the application requirements and the tools to ensure the best network performance.

WAN optimization is about more than just technology. It involves the critical process of aligning network performance with business strategies.

Application Performance Optimization creates an application-oriented WAN infrastructure and helps to achieve a company's network and business goals.

Acronym Key

Term	Explanation
APO	Application Performance Optimization
BOB	Branch Office Box
CDN	Content Delivery Networks
CIFS	Common Internet File System
COS	Class of Service (differentiated services)
FTP	File Transfer Protocol
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
LAN	Local Area Network
MAPI	Messaging Application Programming Interface
OSI	Open Systems Interconnection
QoS	Quality of Service (packet prioritization)
SLA	Service Level Agreement
SMB	Server Message Block
TCP	Transmission Control Protocol
TFTP	Trivial File Transfer Protocol
UDP	User Datagram Protocol
VPN	Virtual Private Network
WAFS	Wide Area File Service
WAN bandwidth	The rate of transmitting data
WAN	Wide-Area Network
WOC	WAN Optimization Controller

Contacts

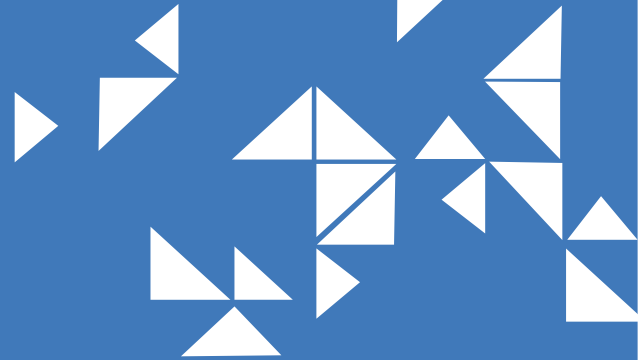
For further information regarding this document or the Tata Communications solutions:

Marketing contact:

Louis Faraud, Tata Communications/, louis.Faraud@tatacommunications.com

Technical contact:

Simvar Lach,Tata Communications/, simvar.lach@tatacommunications.com



Tata Communications, a member of the \$62.5 billion Tata Group, is a leading global provider of a new world of communications. The emerging markets communications leader leverages its advanced solutions capabilities and domain expertise across its global and pan-India network to deliver managed solutions to multi-national and Indian enterprises, service providers and Indian consumers.

Tata Communications' range of services include transmission, IP, converged voice, mobility, managed network connectivity, hosting and storage, managed security, managed collaboration and business transformation for global enterprises and service providers, as well as Internet, retail broadband and content services for Indian consumers.

The Tata Global Network encompasses one of the most advanced and largest submarine cable networks, a Tier-1 IP network, with connectivity to more than 200 countries across 300 Pops, and more than 1 million square feet of data center and co-location facilities.

Tata Communications' unique emerging market depth and breadth of reach includes a national fiber backbone network and access to network in over 60 cities and 125 Pops in India, strategic investments in South African converged services operator, Neotel, Sri Lanka and Nepal and, subject to fulfillment of conditions precedent, a 50% ownership in China Enterprise Communications (CEC) providing full country VPN coverage in China.

Servicing customers from its offices in over 80 cities in 40 countries, Tata Communications is the number one global international wholesale voice operator and number one provider of international long distance, enterprise data and Internet services in India, the Company was named "Best Wholesale Carrier" at the World Communications Awards in 2006, "BestPan-Asian Wholesale Provider" at the 2006 and 2007 Capacity Magazine Global Wholesale Telecommunications Awards and was awarded "Best Progress in Emerging Markets" at the 2008 Mobile Communication Awards.

Tata Communications Limited along with its global subsidiaries, (Tata Communications), is listed on the Bombay Stock Exchange and the National Stock Exchange of India and its ADRs are listed on the New York Stock Exchange. (NYSE: TCL)

www.tatacommunications.com

Tata Communications
"Application Performance Optimization: Challenges and Solutions"