IT@Intel White Paper
Intel IT
IT Best Practices
Network and IT Operations
January 2012

Intelligent WAN Capacity Management at Intel

"As silicon designs become more complex and Intel's product portfolio increases, Intel IT's WAN capacity management solution enables us to optimize network infrastructure investments to meet customer requirements."

Chandra Chitneni
Senior Network Engineer, Intel IT

Chris Clauson
WAN Capacity Manager, Intel IT

Dick Freeman
Senior Network Engineer, Intel IT

Timothy Verrall
Principal Engineer, Intel IT

Executive Overview

Intel IT has developed a highly successful, multifaceted approach to wide area network (WAN) capacity management that enables us to control costs while supporting customer needs and to grow WAN capacity when necessary.

Intel's large and complex WAN interconnects 200 sites around the globe. Increasing Internet usage, complexity of silicon designs, and the breadth of Intel's product portfolio, along with productivity initiatives such as the enterprise private cloud and device-independent mobility, fuel an average annual WAN capacity growth of 10 percent. We estimate that our WAN capacity management process reduces the network costs of international and edge sites by approximately 23 percent per year.

Our WAN capacity management solution comprises four main areas of activity:

- **Infrastructure design and technology.** Intel IT's robust and resilient network design, including Multiprotocol Label Switching, serves as the foundation for an effective WAN capacity management process. We have also implemented a hierarchical quality of service that differentiates between types of network traffic and optimizes the design of Intel's WAN to run as efficiently as possible.
- **Service-level agreements (SLAs).** We use tiered, customer-oriented SLAs, integrated with our three operational-level agreements: design standard, incident escalation, and carrier contracts.
- **Network management tools.** We customized a third-party management tool for use in troubleshooting and forecasting.
- **Formal WAN capacity forecasting process.** We combine data from our network management tool with customer input to forecast future bandwidth requirements. In all cases, we base WAN capacity planning on bandwidth requirements and utilization trends, so that a site's connection is sized properly in the future.

As silicon designs become more complex and Intel's product portfolio increases, Intel IT's WAN capacity management solution enables us to optimize network infrastructure investments to meet customer requirements.

Contents

Executive Overview 1
Background 2
Managing Intel's WAN Capacity 3
- Infrastructure Design and Technology 3
- Service-Level Agreements 4
- Network Management Tools 6
- Formal WAN Capacity Forecasting Process 7

Future WAN Capacity Considerations 7
Conclusion 8
Acronyms 8

IT@INTEL

The IT@Intel program connects IT professionals around the world with their peers inside our organization – sharing lessons learned, methods and strategies. Our goal is simple: Share Intel IT best practices that create business value and make IT a competitive advantage. Visit us today at www.intel.com/IT or contact your local Intel representative if you'd like to learn more.

BACKGROUND

Intel's wide area network (WAN) connects more than 200 sites located in 66 countries, as shown in Figure 1. All sites are connected with the central network; some sites are also connected directly to other sites.

With such a large and interrelated WAN, costs can potentially escalate quickly due to business growth and new acquisitions, as well as varying costs for bandwidth in different geographical regions. We estimate that Intel's WAN capacity grows an average of 10 percent annually,[1] although capacity increases vary in each region of operation. It is critical that we proactively manage Intel's WAN capacity to minimize cost increases.

Changing business requirements and new usage models stimulate the need for additional WAN capacity. For example, while online business tools and strategies reduce total costs for Intel overall, they increase WAN network costs.

[1] Assumption based on post-U.S. core redesign, which added 100 gigabits per second to the U.S. network backbone at reduced overall cost.

- Business groups increasingly use social media to conduct business.
- New applications, such as virtual tape libraries for backup, increase network traffic.
- Consolidation of data centers requires WAN access to remote servers.
- Increased use of videoconferencing and video training materials leads to higher levels of video traffic, which is bandwidth-intensive.
- Applications that display rich visual interpretations of raw data accessed across the network increase data volumes and the need for reduced latency.
- Business continuity and geodiversity result in moving larger amounts of data across the WAN.

As silicon designs become larger and more complex, the data associated with them is also growing, which contributes to WAN traffic. Design teams are distributed across the globe, and the data they create and share needs to move rapidly across continents. Our productivity initiatives such as the enterprise

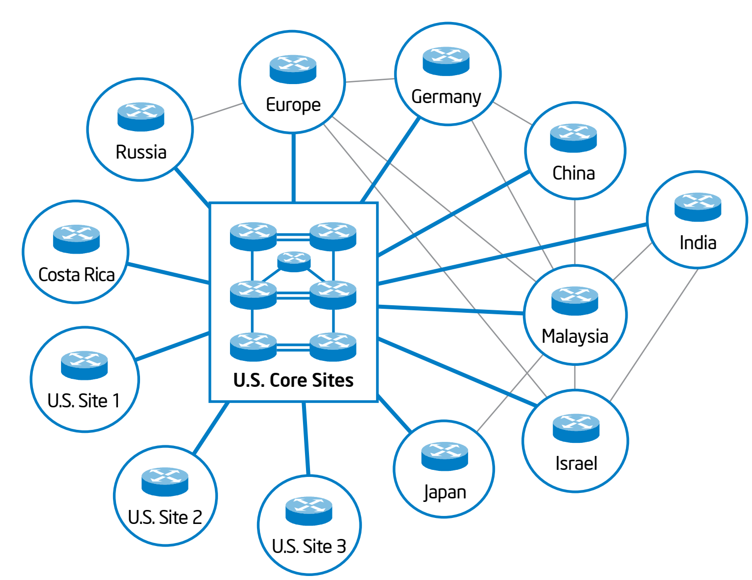

Figure 1. Intel's WAN connects more than 200 sites worldwide.

private cloud and device-independent mobility are also contributing to increasing WAN usage.

With these factors in mind, we've found that managing WAN capacity is the most efficient and cost-effective way to meet our customer's needs.

This approach helps us forecast these needs so bandwidth is available at the right time. It also helps us avoid the costly over-provisioning of bandwidth, which affects our ability to meet customers' needs in a timely manner and ultimately negatively affects our time to market.

MANAGING INTEL'S WAN CAPACITY

Our multifaceted approach to WAN capacity management consists of four core areas of activity: Infrastructure design and technology, service-level agreements (SLAs), network management tools, and a formal WAN capacity forecasting process.

As our capacity management process has evolved over the years, we have found these four areas to be fundamental to implementing an effective WAN capacity methodology. We also found that the successful implementation of these elements requires enhanced collaboration between WAN Capacity Management and WAN Engineering.

Infrastructure Design and Technology

Intel IT's robust and resilient network design serves as the foundation for an effective WAN capacity management process. The key technologies we use include Multiprotocol Label Switching (MPLS) and hierarchical quality of service (QoS). We also optimized the design of Intel's WAN to run as efficiently as possible.

MULTIPROTOCOL LABEL SWITCHING

MPLS technology establishes connectivity between sites using the private Internet Protocol (IP) backbone infrastructure of various regional and global providers. MPLS has key capabilities that deliver tremendous cost advantages:

- Asymmetrical bandwidth. Bandwidth usage tends to be higher from hub to spoke. Instead of paying for the same amount of bandwidth for uplinks and downlinks—which usually requires over-provisioning in at least one direction—the MPLS network enables us to purchase different amounts of incoming and outgoing bandwidth. We pay for the access circuit, downlink speed, and uplink speed separately. With this approach we can purchase the right amount of bandwidth in each direction.
- Large access circuits. We choose access circuits that meet our current needs and have the capacity to accommodate future growth. If we need only a portion of an available access circuit, we provision at the sub-rated level. Using only a portion of the entire large access circuit avoids the cost of unnecessary access circuit usage and ingress bandwidth.
- Agile provisioning. The MPLS network lets us cost effectively provide bandwidth where and when we need it and make adjustments to ingress and egress bandwidth with ease. We provide high-capacity bandwidth for local access; for the remaining distance between two sites, bandwidth increases only when needed.
- Flexible connectivity. Similar to the Internet, our MPLS network enables us to easily connect any two sites, without having to wait for provisioning from the carrier. We can also establish IP connectivity to external partners when needed.

DYNAMIC BANDWIDTH MANAGEMENT USING HIERARCHICAL QUALITY OF SERVICE

Dynamic WAN bandwidth management enables the WAN to deliver a minimum guaranteed bandwidth to each individual site to support proper functionality of voice, video, and other time-sensitive applications. It also enables bursting capability so that any available bandwidth can be made available to the sites that need it, based on relative priority.

Different types of network traffic are more sensitive to delay than others. For example, backing up a database can take 10 minutes or an hour, but the end result is the same. In contrast, a voice packet cannot be half a second late; otherwise, the degradation of quality becomes unacceptable.

Using dynamic bandwidth management and a three-tier QoS policy, we can identify different types of traffic such as video, Internet, database, backup, and voice. Each type of traffic receives a different priority and service guarantee, which we specify in our SLAs.

- Top-tier policy. Matches the egress bandwidth rate to the committed information rate (CIR) purchased from the carrier.
- Second-tier policy. Manages the bandwidth among the multiple tunnels to various Intel sites. The main goal of this policy is to provide guaranteed minimum bandwidth for each site while enabling bursting capabilities to the sites that need the bandwidth the most.
- Third-tier policy. Enables appropriate sharing of bandwidth among various application types (voice, video, time-sensitive, Internet, network backups, and normal traffic) to meet the SLAs for the respective categories.

If network traffic is low, even low-priority data can utilize all available bandwidth. But under load conditions, the bandwidth for each data type may be reduced to its guaranteed level, and the available bandwidth is shared according to the SLA. This approach means business-critical traffic, such as voice and video, receives priority.

By reducing the priority of less critical or time-sensitive traffic, more bandwidth becomes available for applications that need it. At times, however, tiered QoS can increase the volume of traffic at a particular site. "Network Management Tools" discusses how we can analyze these traffic bursts to choose the correct router hardware.

A tiered QoS approach offers several benefits.

- **More efficient bandwidth use across the MPLS links.** Customers can benefit from usage lulls in other locations. For example, design engineers around the world synchronize files daily. Before implementing the three-tiered QoS, synchronization took one hour; after deploying the three-tiered QoS, synchronization took only 19 minutes—at no additional cost.
- **Simplified capacity planning process.** We no longer need to right-size the bandwidth for each tunnel continuously.
- **Reduced costs.** We no longer need to provision each site's CIR equal to the burst need.

Figure 2 shows how the different classes of network traffic are marked when they leave the local area network (LAN) and enter the WAN. The Marking Policy separates traffic on a per-class basis into and across the WAN. Bandwidth is not pre-assigned to queues; QoS queuing is in effect only when congestion occurs. Tiered QoS for routing network traffic gives business-critical traffic, such as voice and video, priority over less time-sensitive applications such as Internet access and network backups. Voice over IP (VoIP) receives a maximum of 66 percent of the bandwidth, while the remaining classes receive a percentage of the residual bandwidth.

WAN OPTIMIZATION

WAN optimization involves traffic deduplication, content caching, and data compression across WAN links to reduce traffic volume. It also improves application performance by supporting local acknowledgements to avoid chattiness across the WAN links and to take advantage of the high-performance Transmission Control Protocol stack for faster transfers across the WAN links.

Optimization delivers LAN-type performance across the WAN, which is essential for certain reporting tools that periodically rely on large data transfers. For situations with repetitive data transfers, only the delta between an already transferred data set and a new version is copied across the WAN.

Service-Level Agreements

We have changed our SLA implementation process. Prior to 2008, we committed to a global SLA of 99.95 percent availability, regardless of customer needs. SLAs did not include committed mean time to repair (MTTR), performance goals, or utilization threshold triggers for adding bandwidth, and they were not aligned with our internal processes.

In 2008, we introduced service-level guidelines that define how we design and manage WAN capacity for a particular site, based on customer needs at that site. We use a tiered SLA structure based on site classifications, which are primarily based on a site's criticality to Intel's business. We commit to meeting the SLAs across site tiers, not individual sites, and we review our SLA performance regularly with our customers.

Customer-oriented SLAs help us determine how a site should be connected and generally include a number of key components that contribute to effective WAN capacity management.

- The utilization threshold that will trigger an upgrade
- A specification for the WAN redundancy (required number of routers and circuits)
- Failover capacity
- The services to be supported, such as distributed design computing, webcast,

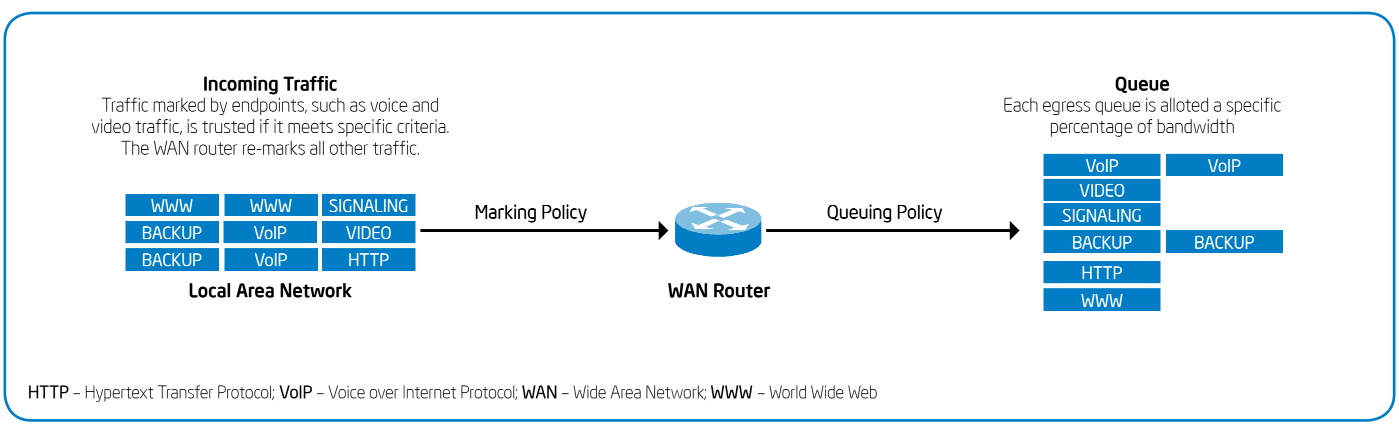

Figure 2. We use a tiered quality of service approach for routing network traffic.

VoIP videoconferencing, internal and external cloud computing, and virtual desktop infrastructure (VDI)

- Committed availability and MTTR requirements

SLAs are integrated with our three operational-level agreements (OLAs): design standard, incident escalation, and carrier contracts.

DESIGN STANDARD OLA

Table 1 lists the design standards for three site tiers with varying router and circuit configurations, levels of failover capacity, and utilization thresholds. We make exceptions to these recommendations, as necessary, to satisfy specific business requirements.

INCIDENT ESCALATION OLA

Not all incidents require equal responses. We tier our MTTR commitments according to the severity of the incident and the site classification.

Figure 3 shows how we correlate the site tier with the severity of the incident to define four levels of MTTR. In determining the severity of an incident, we consider how much network performance is degraded, whether the degradation is isolated to a particular site, and the time at which the incident occurs. We assign a higher severity to incidents that occur during standard business hours—roughly 7:00-18:00 local time.

Carrier Contracts OLA

We review our WAN carrier contracts proactively to identify and address gaps between carrier SLAs and our customer-committed SLAs. For example, a customer-committed SLA might stipulate a one-hour MTTR but the carrier's SLA may indicate a two- to four-hour MTTR.

We also include standard terms and conditions in carrier SLAs, and we review performance against carrier SLAs quarterly.

Service Availability

Using our SLAs and OLAs, we commit to a specific level of service across each tier, not for each individual site. Table 2 lists service availability according to SLA level and the type of site.

Table 1. Design standard recommendations

	Router or Circuit	Diverse Central Office An Intel site that connects to two different carrier "central office" locations	Failover Capacity	Utilization	Internal Intel Telephony Dialing Plan	Multicast
Tier 1 Platinum	Dual	Yes	100%	50%	Yes	Yes
Tier 2 Platinum	Dual	Yes	50%	70%	Yes	Yes
Tier 3 Platinum	Dual	Yes	100%	50%	Yes	Yes
Tier 3 Gold	Dual	Recommended	50%	80%	Yes	Yes
Tier 3 Silver	Single	No	NA	90%	No	Yes
Tier 3 Bronze	VPN	NA (Not Applicable)	NA	100%	NA	Yes

VPN –Virtual Private Network

Operational-level Agreements Escalation

IMPACT DESCRIPTION	URGENCY CATEGORIES: CRITICAL	HIGH	MEDIUM	LOW
Extensive and Widespread Tier 1, Tier 2, Tier 3 Platinum	▪ Isolated	▪ Severe performance degradation	▪ Minor performance degradation includes voice ▪ Loss of bandwidth with no impact	▪ Minor error detected
Significant and Large Tier 3 Gold	Not Applicable (NA)	▪ Severe performance degradation ▪ Isolation during prime time	▪ Minor performance degradation includes voice ▪ Isolation during non-prime time	▪ Minor error detected ▪ Loss of bandwidth with no impact
Moderate and Limited Tier 3 Silver	NA	▪ Severe performance degradation ▪ Isolation during prime time	▪ Minor performance degradation includes voice ▪ Isolation during non-prime time	▪ Minor error detected ▪ Loss of bandwidth with no impact
Minor and Localized Tier 3 Bronze	NA	▪ Severe performance degradation ▪ Isolation during prime time	▪ Minor performance degradation includes voice ▪ Isolation during non-prime time	▪ Minor error detected ▪ Loss of bandwidth with no impact

MEAN TIME TO REPAIR
- 1 hour
- 4 hours
- 8 hours
- 48 hours (consultation)

Figure 3. Different tiers of service included in an operational-level agreement categorize the urgency of a network issue, how far-reaching its impact is, and the mean time to repair for each tier.

Table 2. Service availability as defined by wide area network site service-level agreement tier and type

	Site Type	Service Availability
Tier 1 Platinum	U.S. Core	99.99%
Tier 2 Platinum	Hub or aggregation	99.99%
Tier 3 Platinum	Critical manufacturing and logistics facilities	99.99%
Tier 3 Gold	Key design and SMG facilities	99.99%
Tier 3 Silver	Small-to-medium field sales office facilities	99.99%
Tier 3 Bronze	Small field sales offices	99.99%

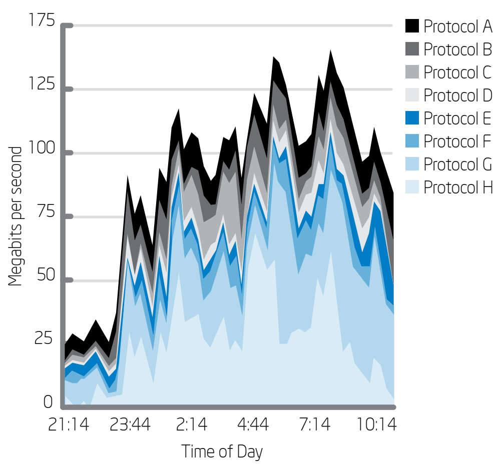

Figure 4. Using a network management tool, we can track the flow of packets across the network for any connection.

Network Management Tools

We customized a third-party management tool that we use for troubleshooting and forecasting. The tool records network utilization data on every WAN circuit globally to provide statistics and graphical views detailing the history of our WAN links. We analyze this history to forecast future bandwidth needs over a 12- to 18-month period. The tool also provides application-specific data that helps us comprehend the type of traffic going across the link.

We cannot make informed capacity upgrade decisions by simply monitoring the raw utilization of a circuit. Therefore, instead of combining all network traffic across a circuit, we differentiate application traffic into different queues and monitor the utilization at a per-queue level. All network backup traffic uses one queue and higher utilization of this queue may not drive capacity upgrades. Similarly, there are different queues for voice, video, and time-sensitive traffic; the criteria for managing bandwidth is different for each of these queues.

Several features of the network management tool are crucial to forecasting and troubleshooting.

SYNTHETIC TRANSACTIONS

We can test transaction performance across the network without actually sending any user data. For example, we can test database queries, simulate video and voice calls, and measure the performance levels users would experience.

NETWORK FLOW DATA

The network management tool records the flow of packets across the network for any connection. Data includes origin at the machine level, flow direction, port numbers, duration of the transaction, total traffic volume, and the application associated with each packet. This tool provides central access to this data in a transaction log as well as graphical views. We use the network flow data to help forecast bandwidth requirements and to help network engineers diagnose network problems. Figure 4 shows a sample of network flow data. Each color in the graph represents a different protocol.

QUALITY OF SERVICE (QoS)

We monitor bandwidth usage and packet drops for each of the class-based queues for all the WAN connections. This enables us to monitor the bandwidth usage for various application types and make adjustments to the QoS policy based on actual usage. Figure 5 shows a screenshot of the QoS data that our network management tool provides.

From the QoS data, we can determine the data transfer rate before and after queueing and track how many packets are dropped. The goal for higher-tier queues is to have zero dropped packets—that is, pre-queue volume and post-queue volume are equal. For lower-tier queues, such as the "out-best-effort" queue, some dropped packets are acceptable.

Name	Pre Packets	Pre Volume	Pre Bit Rate	Post Volume	Post Bit Rate
class-default	38.01 M	18.58 GB	42.69 Mbps	18.58 GB	42.69 Mbps
out-best-effort	24.51 M	24.21 GB	54.54 Mbps	24.21 GB	54.54 Mbps
out-voice	8.53 M	1.72 GB	3.76 Mbps	1.72 GB	3.76 Mbps
out-proxy	4.25 M	3.45 GB	6.42 Mbps	3.45 GB	6.42 Mbps
out-streaming	1.58 M	1.05 GB	2.41 Mbps	1.05 GB	2.41 Mbps
out-virus-control	1.54 M	137.49 MB	301.67 Kbps	137.49 MB	301.67 Kbps
out-time-sensitive	360.55 K	65.68 MB	125.33 Kbps	65.68 MB	125.33 Kbps
out-sync-voice	5.59 K	2.01 MB	6.58 Kbps	2.01 MB	6.58 Kbps
out-troubleshooting	0	0 B	0 bps	0 B	0 bps

Figure 5. We also use our network management tool to monitor bandwidth usage and packet drops for all queues.

SIMPLE NETWORK MANAGEMENT PROTOCOL QUERIES

We collect utilization and network health statistics from the network hardware, which enables us to accurately determine peak, average, and 95th percentile loads of network circuits and our hardware infrastructure.

ALERTS

We use alerts to proactively manage site performance. Our network management tool generates integrated performance monitoring alerts when a site experiences heavy congestion or when the network performance of a certain queue drops below the performance threshold. We can then analyze traffic patterns at the site and help improve the performance by working directly with the customer or application owner. IP SLA alerts that are ongoing may require a bandwidth upgrade.

Formal WAN Capacity Forecasting Process

We combine data from our network management tool with customer input to forecast future bandwidth requirements. In all cases, we base WAN capacity planning on bandwidth requirements and utilization trends, so that a site's connection is sized properly in the future. We use the following activities to help us forecast WAN capacity:

- A yearly financial exercise that results in an investment roadmap that support business needs for the upcoming year
- A formal customer engagement process to identify planned changes and assess future requirements in different regions and sites
- Automated business intelligence to track utilization trends

Before implementing our network management tool, we based forecasts on overall network utilization, regardless of traffic type. During times of peak bandwidth usage, we couldn't determine whether the problem was due to backup traffic or something more business-critical, resulting in our upgrading the capacity.

With our new network management tool in place, we can differentiate among traffic types, and purchase bandwidth more wisely. For example, backup traffic has a lower priority than interactive and silicon design traffic. If a utilization spike is due to backup traffic only, we will not actively invest in additional bandwidth. If interactive silicon design applications or videoconferencing are causing the spike, we may decide to invest in an upgrade for a particular site because these types of traffic are more business-critical.

Figure 6 shows a sample of the data we analyze to determine current utilization and predict when we will reach peak capacity.

FUTURE WAN CAPACITY CONSIDERATIONS

Intel's WAN capacity will continue to grow as Intel's product portfolio increases, product designs become more complex, and productivity initiatives such as the enterprise private cloud and device-independent mobility become even more prevalent. Some of the specific areas we are watching include:

- Increasing number of video sites, services, and solutions
- The need for lower latency connections in support of increasing interactive and real-time collaboration requirements, and improved throughput for site-to-site file transfers
- Increasing volumes of backup traffic as Intel's employee base and hard disk space utilization grows

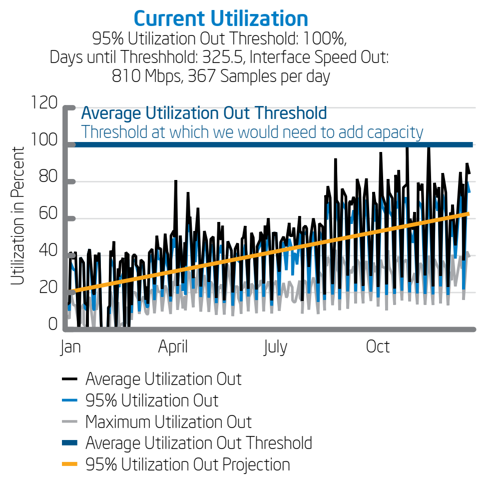

Figure 6. Analyzing current utilization helps us predict when we will reach peak capacity and plan for adding capacity only when it is needed.

- External cloud applications and services, which place a heavy load on Internet connections and impact WAN links to remote sites
- Virtual conferences, which can place a massive load on Internet connections—as much as 65 kilobits per second (Kbps) to 500 Kbps, per user. For example, Intel's virtual 2010 International Sales and Marketing Conference had close to 5,000 attendees. Figuring a maximum of 500 Kbps per user, that is a total of 2.5 gigabits per second.
- Dynamic call admission control, which may be using resource reservation protocol (RSVP) for all real-time applications, such as voice and video. RSVP is used to "guarantee" bandwidth to specific applications and sessions. This can increase bandwidth demand in certain situations, but can also lower demand if the design is set up correctly.
- Prioritization of Design Engineering applications to ensure that these applications get the critical performance and throughput they require.

CONCLUSION

Increasing Internet usage, product complexity, and product portfolio fuel an annual average WAN capacity growth of 10 percent. We can control network costs while still meeting customer bandwidth requirements by focusing on four areas of WAN capacity management.

- Infrastructure design and technology for a robust, resilient, and cost-effective network
- Tiered SLAs that prioritize different types of network traffic
- Network management tools that support troubleshooting and forecasting
- A formal WAN capacity forecasting process to help us accurately predict future bandwidth needs and grow our WAN capacity when necessary

We estimate that our highly successful, multifaceted WAN capacity management process saves Intel approximately 23 percent per year for our international and edge site connectivity.

ACRONYMS

CIR	committed information rate
IP	Internet Protocol
Kbps	kilobits per second
LAN	local area network
MPLS	Multiprotocol Label Switching
MTTR	mean time to repair
OLA	operational-level agreement
QoS	quality of service
RSVP	Resource Reservation Protocol
SLA	service-level agreement
VoIP	Voice over Internet Protocol
WAN	wide area network

For more information on Intel IT best practices, visit www.intel.com/it.

This paper is for informational purposes only. THIS DOCUMENT IS PROVIDED "AS IS" WITH NO WARRANTIES WHATSOEVER, INCLUDING ANY WARRANTY OF MERCHANTABILITY, NONINFRINGEMENT, FITNESS FOR ANY PARTICULAR PURPOSE, OR ANY WARRANTY OTHERWISE ARISING OUT OF ANY PROPOSAL, SPECIFICATION OR SAMPLE. Intel disclaims all liability, including liability for infringement of any patent, copyright, or other intellectual property rights, relating to use of information in this specification. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted herein.

Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and other countries. * Other names and brands may be claimed as the property of others.

Copyright © 2012 Intel Corporation. All rights reserved. Printed in USA ♻ Please Recycle 0112/JGLU/KC/PDF 326383-001US