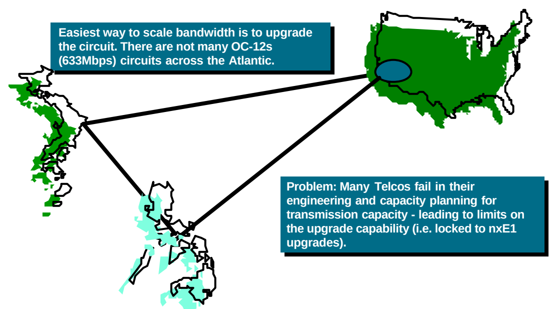
Trans-Oceanic Backbones Basic Techniques

- Six Techniques used by the Internet Community to scale bandwidth:
  - ✓ Bigger Circuits
  - ✓ Inverse Multiplexing
  - ✓ Clear Channel E3 or DS3
  - ✓ PPP over SDH
  - ✓ Asymmetrical Satellite Systems
  - ✓ Hybrid Systems

Trans-Oceanic Backbones Basic Techniques

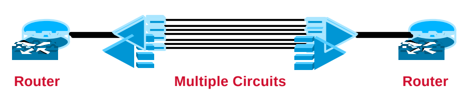


Trans-Oceanic Backbones Basic Techniques

> Pearl: It helps to work directly with the people who do the international transmission capacity planning and purchasing. They get to see your projections, you get on time upgrades.

Trans-Oceanic Backbones Basic Techniques

- Inverse Multiplexing (iMux) takes several circuits and *bundles* them into one or more logical circuits.
- Two major techniques:
  - ✓ Use protocol/forwarding features in the router
  - ✓ Use an external inverse multiplexer



Trans-Oceanic Backbones
Basic Techniques

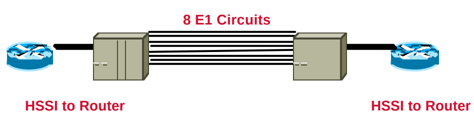
- Router protocol/forwarding features as iMux
  - ✓ Parallel Links Across the Ocean. n x E1 circuits between the routers using the routing protocols to perform the load balancing and bundling of the parallel circuits - works up to 4 to 8 E1s.

Trans-Oceanic Backbones
Basic Techniques

- Several Techniques:
  - ✓ Static Route - Per Packet
  - ✓ OSPF
  - ✓ eBGP Multihop - Per Flow (w/ Netflow & CEF) or Per Packet (w/ CEF)
  - ✓ Multi-Link PPP (MLPPP) - Tighter Bundling Options (up to 8)
  - ✓ CEF Load Balancing - Per packet or per flow
  - ✓ BGP Maximum Paths (up to 6 - different routers)

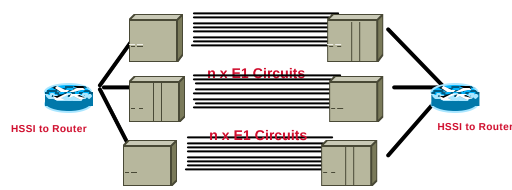
Trans-Oceanic Backbones
Basic Techniques

- ✓ Many Telcos have not provisioned facilities to cater to oceanic circuits above E1.
- ✓ Many E1s grouped together into a larger pipe via iMUX technology.
- ✓ Defacto Industry practice is to use Cisco Routers with HSSI ports connected to Larscom's Orion 4000 iMUXes.



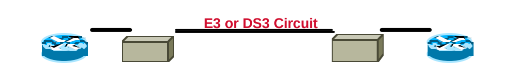
Trans-Oceanic Backbones
Basic Techniques

- ✓ Several iMUX bundles can be grouped together on the same router to build 34M and 45M equivalent circuits
- ✓ eBGP Multihop is the preferred load balancing technique.
- ✓ Telstra Internet is now over 100M of iMUX backbone bandwidth!



Trans-Oceanic Backbones
Basic Techniques

- Clear Channel E3 (34M) or DS3 (45M)
  - ✓ Preferred method for high speed backbone links is a clear channel circuit.
  - ✓ Configuration is simple - connect the CSU/DSU to a HSSI/T!/E1 port on a Cisco router.

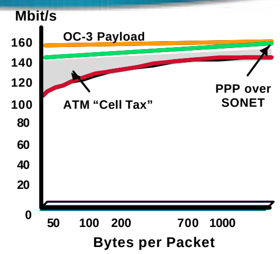


Trans-Oceanic Internet
Backbones

- What do you do after DS3 (45M)?
  - ✓ Multiple DS3?
  - ✓ ATM at OC-3 (155M) or OC-12 (622M)?
  - ✓ PPP over SDH at OC-3 (155M) or OC-12 (622M)?
- ATM vs PPP over SDH (POSIP)
  - ✓ ATM is <u>not</u> the best choice when all you are doing is Internet traffic across the ocean.

Trans-Oceanic Internet Backbones

- **Line utilization**
  - ✓ Packet over SONET provides 98% utilization
  - ✓ 20% - 30% overhead over an ocean is a lot of money!
- **Goodput verses throughput**
  - ✓ Sufficient buffering for large TCP flows
  - ✓ Congestion avoidance
  - ✓ ATM and TCP/IP Headers are part of "throughput," not "goodput"



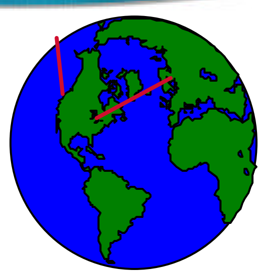
Trans-Oceanic Backbones Basic Techniques

POSIP Overview

- ✓ Packet-oriented serial interface, OC3/STM-1, OC-12, OC-48
- ✓ Supports either SONET or SDH interface
- ✓ PPP packets are encapsulated in SONET STS-3c
- ✓ Provides superior line utilization and data efficiency
  - POS available information bandwidth: 149.76 Mbps
  - ATM available information bandwidth: 128.36 Mbps
  - Cell tax ranges from 14-50% depends on the packet distribution
- ✓ Encapsulations
  - RFC 1619 point-to-point protocol over SONET/SDH
  - RFC 1662 point-to-point protocol in HDLC-like framing
- ✓ IP protocol only

Trans-Oceanic Internet Backbones

- **Pensacken NY (Sprint NAP) to Stockholm Sweden (D-GIX).**
- **First 155 M Internet link across an ocean!**
- **First operational 155M POSIP connection!**
- **Second is from Japan to US**

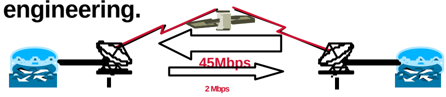


Trans-Oceanic Backbones Basic Techniques

- **If an ISP's trans-oceanic traffic pattern is always asymmetrical, then why pay for that idle bandwidth going out?**
  - ✓ **Asymmetrical traffic pattern is when the ISPs is pulling down more information than sending out. Today a typical ISP is somewhere between a 80:20 - 60:40 traffic ratio to the US.**
- **Satellite Services allow ISPs to buy a circuit with different speeds in the two different directions.**
  - ✓ **For example - 2 Mbps in / 256 Kbps out**
- **Hence the ISPs only pays for what they need - no idle bandwidth giving the other side a free ride.**

Trans-Oceanic Backbones Basic Techniques

- **Trans-Oceanic ISP links are a reality. They work and are used to backup terrestrial trans-oceanic links.**
- **Latency issues (around 700 ms RTT) can be minimized through creative engineering.**

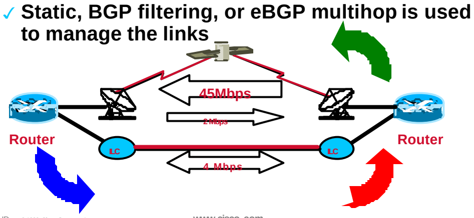


Trans-Oceanic Backbones Basic Techniques

- **Minimizing Latency on Asymmetrical Satellite Links:**
  - ✓ **Good Traffic Engineering (i.e. avoid congestion)**
  - ✓ **Hybrid Asymmetrical Links**
  - ✓ **Hybrid Simplex Links**
  - ✓ **WWW Caching**
  - ✓ **Content Routing**

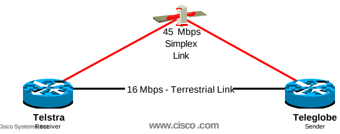
Trans-Oceanic Backbones Basic Techniques

- Hybrid Asymmetrical Satellite links combine terrestrial and satellite together.
  - ✓ Reduces the latency by 1/3 to 1/2.
  - ✓ Static, BGP filtering, or eBGP multihop is used to manage the links



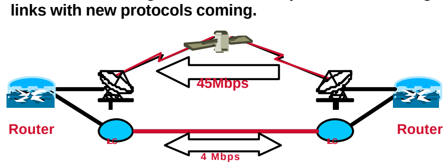
Trans-Oceanic Backbones Basic Techniques

- Telstra and Teleglobe were the first two ISPs who pioneered this technique.
  - ✓ Several other ISPs in Asia and Europe are using this technique.
  - ✓ Very few Tier 1 & 2 NSPs in the US will terminate these trans-oceanic systems - hence the growth of co-lo business (I.e. AboveNet)



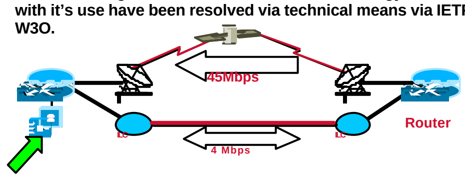
Trans-Oceanic Backbones Basic Techniques

- Hybrid Simplex Satellite links combine terrestrial and satellite's capability to for a circuit with only direction..
  - ✓ Takes advantage of uni-directional nature of satellite circuits
  - ✓ Reduces the latency by 1/3 to 1/2.
  - ✓ Static, BGP filtering, or eBGP multihop is used to manage the links with new protocols coming.



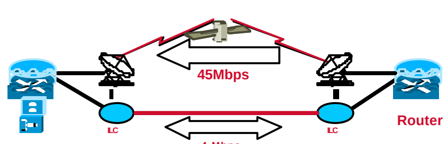
Trans-Oceanic Backbones Basic Techniques

- WWW Caching adds an additional "buffer" to the higher latency.
  - ✓ Caches content as it comes over the link.
  - ✓ Minimizes the download of the same content over and over again.
  - ✓ WWW Caching is a main stream Internet technology - all issues with it's use have been resolved via technical means via IETF and W3O.



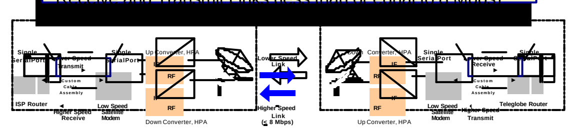
Trans-Oceanic Backbones Basic Techniques

- Since asymmetrical systems will limit any "free rides" from people who should be paying for the traffic from the ISP to the world.
  - ✓ The system is designed around the ISP's traffic profile.
  - ✓ Choke point in the ISP's space limits excessive pull from the other side.



Trans-Oceanic Internet Backbones

Asymmetric Satellite Service - Scenario 1A
ISP Router in Earth Station
Receive and Transmit Links (less than or equal to 8 Mbps)

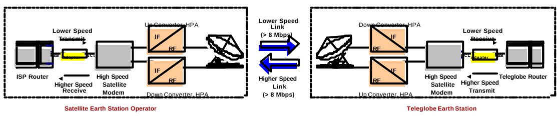


Single low-speed satellite modem supports up-link and down-link speeds up to 8 Mbps

Slides courtesy of Teleglobe

Trans-Oceanic Internet Backbones

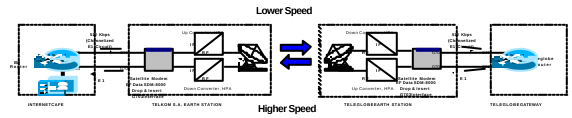
Asymmetric Satellite Service - Scenario 1B
ISP Router in Earth Station
Receive Link to ISP (greater than 8 Mbps)
Transmit Link from ISP (greater than 8 Mbps)



Single high-speed satellite modem supports up-link and down-link speeds up to 45 Mbps

Trans-Oceanic Internet Backbones

Teleglobe & Intelsat Asymmetric E1/512 Kbps Satellite Link
for Africa Telecom `98 Internet Cafe



Trans-Oceanic Internet Backbones

- **Equipment Issues**
  - ✓ **Need a interface card in a router with lots of buffering. VIP2-50 with max memory connected to a HSSI PA or POS PA (DS-3) are known to work.**
  - ✓ **Need Random Early Detection (RED). Needed to insure effective utilization of the link.**

New Trends

- **Co-Location or Lease of Router in the US.**
- **Dual Sided Content Routing.**
- **Heavy Localization of traffic (IXPs)**

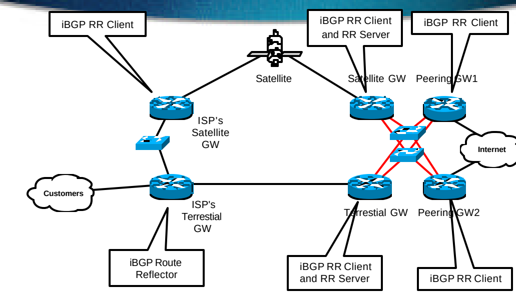
New Trends

- **ISP's Trans-Oceanic Backbones are migrating into *systems* designed to get maximized efficiency from the link.**
- **High Cost of trans-oceanic bandwidth, Exponential Growth, and new demand for Value Added Services (QoS, Content Routing, and VoIP) are all driving factors.**

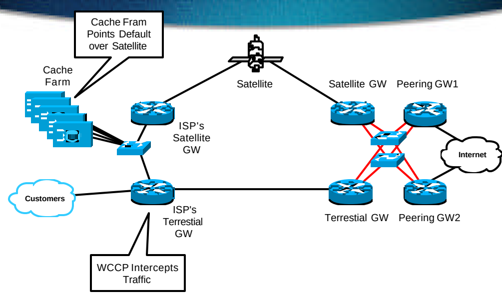
New Trends

- **These Trans-Oceanic Systems will consist of:**
  - ✓ **Mix of Satellite and Terrestrial Circuits**
  - ✓ **WWW Caching and Performance Enchanting Proxies**
  - ✓ **QoS Services (Premium and Basic)**
  - ✓ **Application Redirection (Voice and Multicast)**

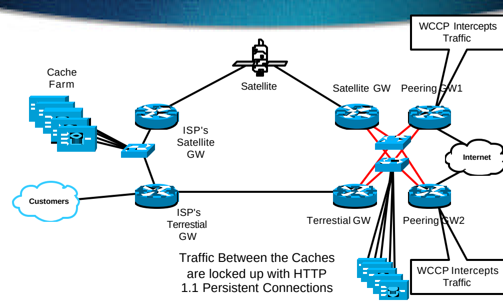
New Trends



New Trends



New Trends

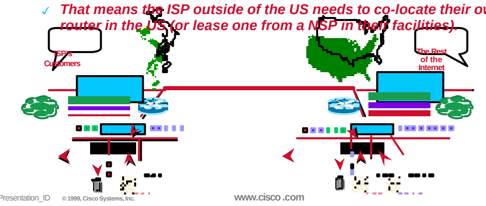


New Trends

- **HTTP 1.1 Persistent connections between the two caches move the average packet size from ~40 bytes to 512 - 1500 bytes.**
  - ✓ **More Goodput vs Overhead**
  - ✓ **Satellite Modifications to TCP increase efficiencies (RFC 2488 or equivalent)**
- **Technique is also called *Performance Enhancing Proxies (PEP)***
  - ✓ `draft-ietf-pilc-pep-00.txt`

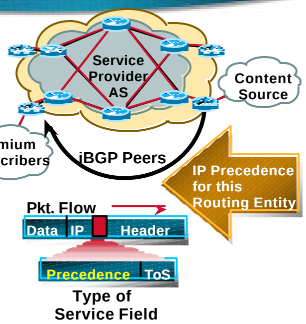
New Trends

- **Any QoS, CoS, or DiffServ tools need to be applied on the *upstream* router's interface.**
  - ✓ **Applying the tools on the *downstream* side would force the ISP to pay for the packets before they are dropped.**
  - ✓ **Most US NSPs are reluctant to apply any special configurations on the US side.**
  - ✓ ***That means the ISP outside of the US needs to co-locate their own router in the US (or lease one from a NSP in their facilities).***

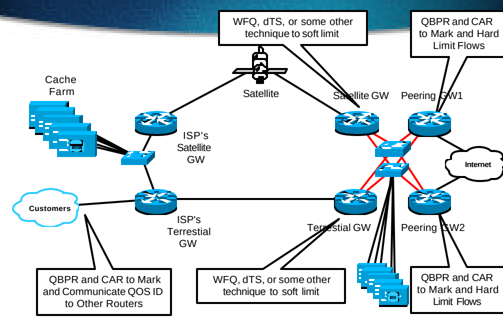


New Trends

- Use BGP to propagate precedence-setting policy as needs change
- BGP community label used to bind precedence polices to destination prefixes
- Return traffic classified with appropriate QoS

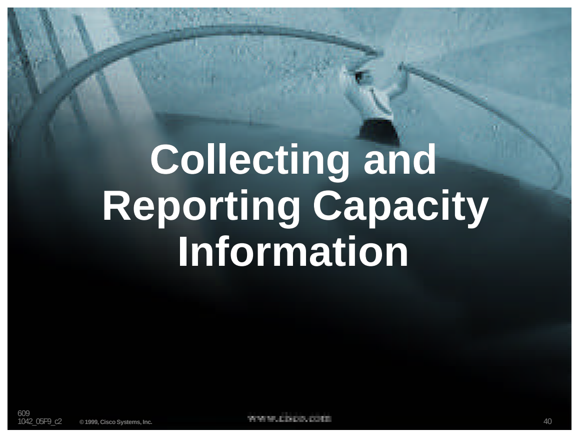


New Trends



New Trends

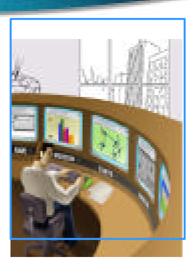
- What's Next?
  - ✓ VoIP and IP Telephony. Using routing tricks or NPR (Netflow Policy Based Routing) to keep voice traffic on the terrestrial circuits.
  - ✓ Content Replication. Pushing content from the international side to the US side of the link.



Collecting and Reporting Capacity Information

Internet Traffic Measurement

- Aggressive collections and analysis of network data is critical to any ISPs who wishes to tackle the problems of CoS, QoS, and I3F
- Ironically, most ISPs do not collect this information, even when most of the tools are public domain on the Internet.
- The concern is that so many people are talking about buzzword and not enough about the fundamentals of what is actually happening on the Internet.



Performance Management and Capacity Planning Definitions

- Capacity planning
  - ✓ The process of determining the likely future network resource requirements to prevent a performance impact on business critical applications
- Performance management
  - ✓ The practice of managing network service response time, consistency and quality for individual services and services overall

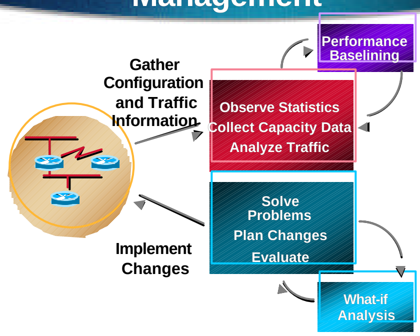
Increasing Importance of Capacity Planning

- Frequent application deployment failure
- Increased reliance on network services for business applications
- Exponential growth in business and nonbusiness related traffic
- Network Failure is typically capacity related

Capacity Related Risks

- Network degradation and failure
- Application timeouts and failure
- Application performance degradation

Effective Capacity Management

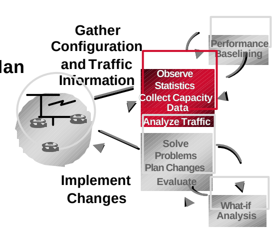


Resource Constraints or Bottlenecks

- CPU
- Memory
- Buffering, queuing and latency
- Interface and pipe sizes
- Speed and distance
- Application characteristics

Collecting and Reporting Capacity Information

- Development of information collection plan
- Tools for collecting capacity information
- Defining capacity areas
- Reporting and interpreting results



Information to Collect

- Link utilization
- CPU
- Memory
- Performance (ping response time)
- Queue/buffer drops
- Broadcast volume
- Frame Relay DE, FECN, BECN, traffic-shaping parameters
- NetFlow statistics
- RMON

Link Utilization

Resource	Address	Segment	Avg. Util (%)	Peak Util (%)
JTKR01S2	10.2.6.1	128 Kbps	66.3	97.6
JYKR01S0	10.2.6.2	128 Kbps	66.3	97.8
FMCR18S4/4	10.2.5.1	384 Kbps	51.3	109.7
PACR01S3/1	10.2.5.2	384 Kbps	51.1	98.4

CPU Utilization

Resource	Polling Address	Avg. Util (%)	Peak Util (%)
FSTR01	10.28.142.1	60.4	80
NERT06	10.170.2.1	47	86
NORR01	10.73.200.1	47	99
RTCR01	10.49.136.1	42	98

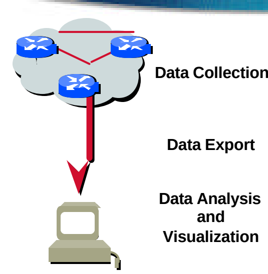
Performance (Ping Response Time)

Resource	Address	AvRes T (mS) 09-09-98	AvRes T (mS) 09-09-98	AvRes T (mS) 09-24-98	AvRes T (mS) 10-01-98
AADR01	10.190.56.1	469.1	852.4	461.1	873.2
ABNR01	10.190.52.1	486.1	869.2	489.5	880.2
APRR01	10.190.54.1	490.7	883.4	485.2	892.5
ASAR01	10.196.170.1	619.6	912.3	613.5	902.2
ASRR01	10.196.178.1	667.7	976.4	655.5	948.6
ASYR01S					503.4
AZWRT01	10.177.32.1	460.1		444.7	
BEJR01	10.195.18.1	1023.7	1064.6	1184	1021.9

Three Essential Tools

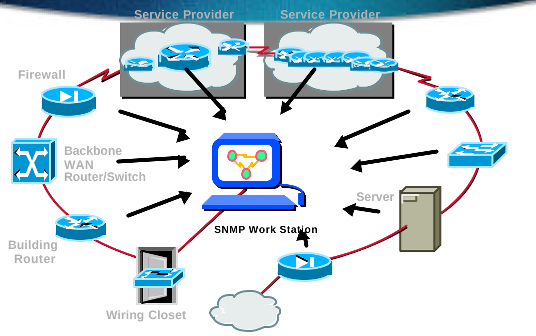
- Simple Network Management Protocol (SNMP)
- Remote MONitoring Protocol (RMON)
- NetFlow - Flow Based TCP/IP Analysis

Traffic Management Elements

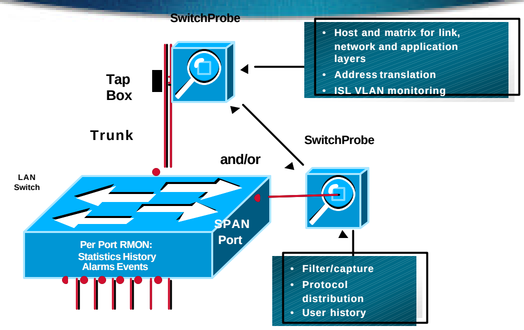


- Data collection mechanisms on network equipment
- Data export mechanisms to applications
- Data analysis and visualization

SNMP is everywhere in the Internet



Network Monitoring with RMON

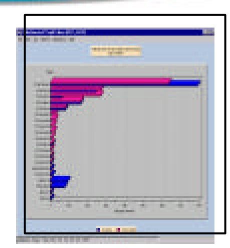


Traffic Analysis on: Link, Network and Application Layers

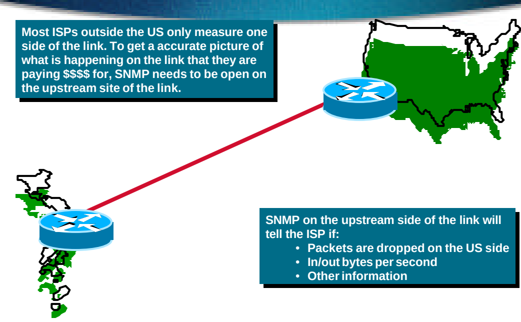
- Aggregate and historical statistical analysis for switched segments
  - ✓ Bandwidth utilization
  - ✓ Error analysis
  - ✓ Broadcast levels
  - ✓ Baseline analysis

Top Hosts

- Top hosts by any of the following metrics:
  - ✓ Packets sent or received
  - ✓ Octets sent or received
  - ✓ Broadcasts sent
  - ✓ Multicasts sent
  - ✓ Errors generated

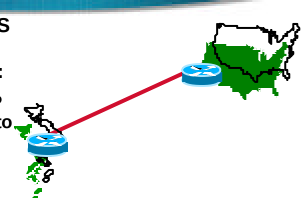


Measuring International Links

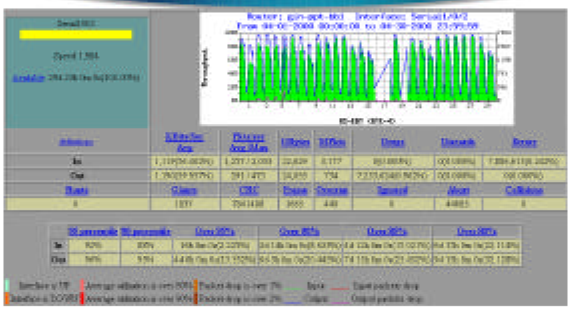


Measuring International Links

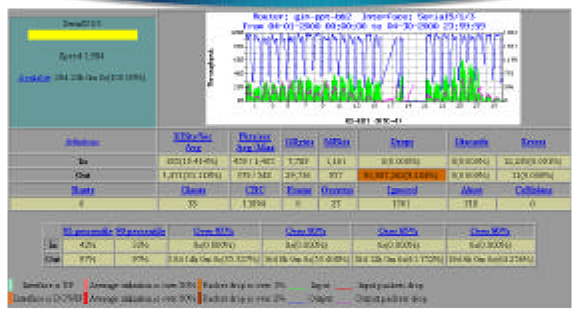
- ISPs Outside of the US should require the upstream provider to:
  - ✓ Create a special SNMP Community with ACL to poll the router on the US side.
- OR
  - ✓ Create a special Web page that displays the SNMP data. MRTG or Looking Glass with access security would work.



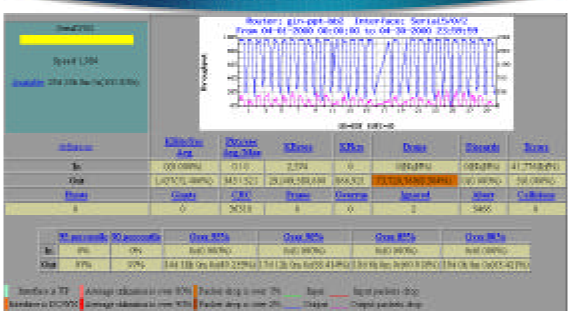
Measuring International Links



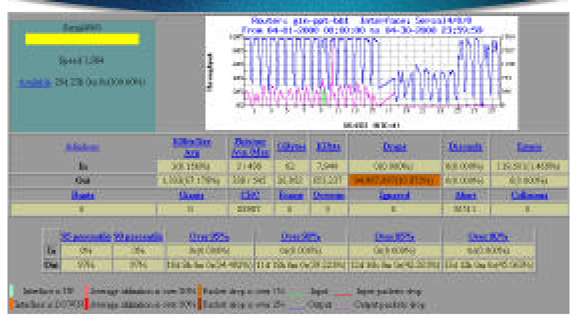
Measuring International Links



Measuring International Links



Measuring International Links

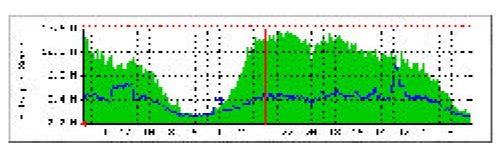


Fundamental Tools are Cheap!

- **The tools to create a simple network manage system that will give an ISP the basics comprise of the following:**
  - ✓ PC with LINUX (free UNIX)
  - ✓ CMU SNMP (free SNMP)
  - ✓ PERL5 (free UNIX script language)
  - ✓ GNU Plot (free graphic plot tool)
  - ✓ Printer

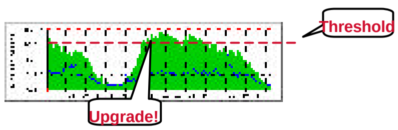
Fundamental Tools are Cheap!

- **Example of what can be done with another Shareware tool - MRTG ….**



Fundamental Tools allow for the Baseline!

- Baseline Quality Levels are critical for any ISP Server.
  - ✓ Average Utilization and Packet Loss need to be monitored on the entire network.
  - ✓ QoS Threshold need to be set and acted on to maintain any sort of foundation to build advanced IP services. This is ***ISP 101*** which most new ISPs forget!
  - ✓ All you need is SNMP! It's not rocket science.

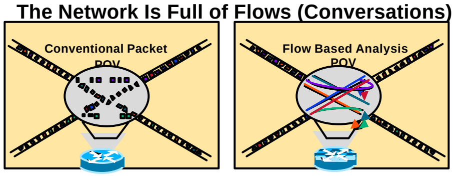


Limitations of SNMP and RMON

- **SNMP and RMON will tell you what is happening on the network (I.e. load, PPS, packet drops).**
- **SNMP and RMON will not tell you <u>who</u> is doing <u>what</u> to <u>where</u> and <u>when</u>.**
- **For that sort of details, *TCP/IP Flow Based Analysis* is needed.**

Flow Based Analysis

Connectionless IP But…
The Network Is Full of Flows (Conversations)



- **Flows are Unidirectional**
- **Flows are Granular**
  - **- IP address and app. port# pairs**
  - **- (TOS/Protocol/Interface)**

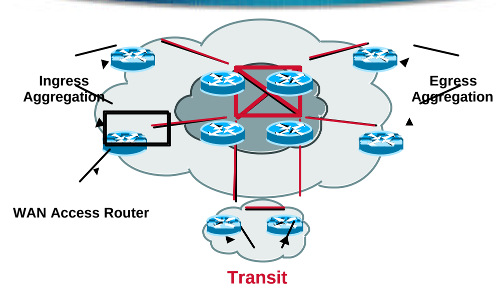
Flow Based Analysis

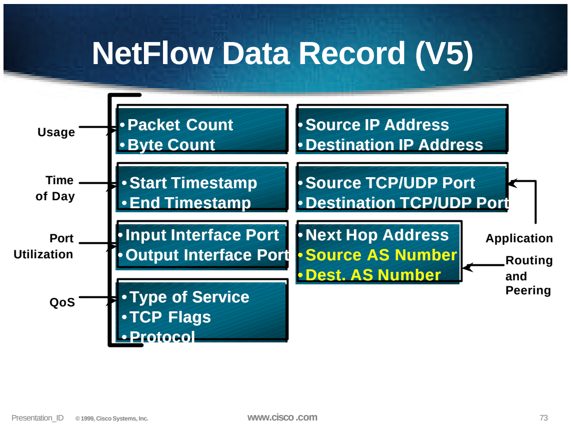
- **Key IETF work:**
  - ✓ **Real Time Traffic Flow (RTFM) working group**
    - ⇨ **http://www.auckland.ac.nz/net/Internet/rtfm/**
  - ✓ **IP Provider Metrics (IPPM)**
- **Public Domain and Commercial Tools now available.**

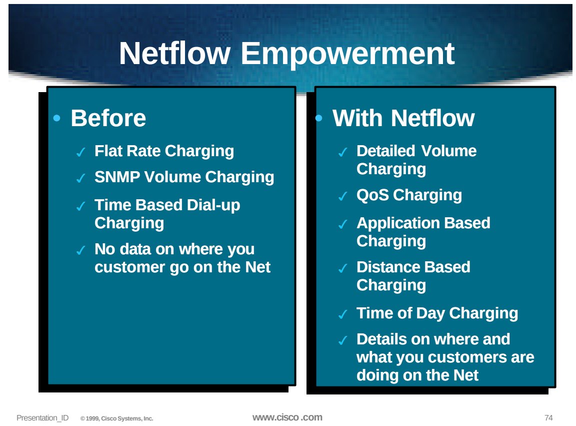
Flow Based Analysis

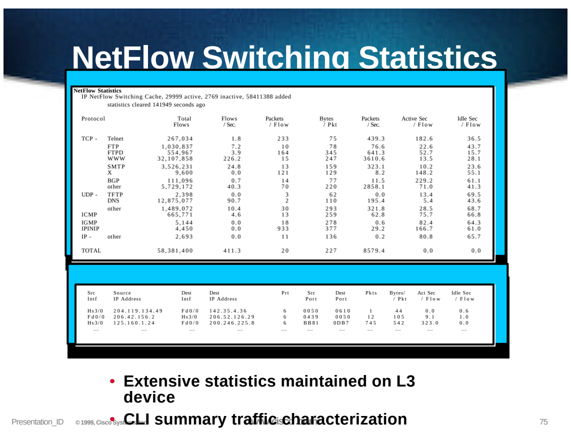
- **Key Tools Used Today**
  - ⇨ **NetScarf** - Global and Regional Internet Analysis project (http://www.merit.edu/~netscarf)
  - ⇨ **Traffic Flow Analysis** - NLANR (http://www.nlanr.net)
  - ⇨ **NetFlow** - Analysis and IP switching technology build into Cisco's IOS.
  - ⇨ **NeTraMet** - Free Flow analysis software. (http://www.auckland.ac.nz/net/Accounting/ntm.Release.note.html)
  - ⇨ Many new Netflow based commercial tools ….

Flow Activation Locations



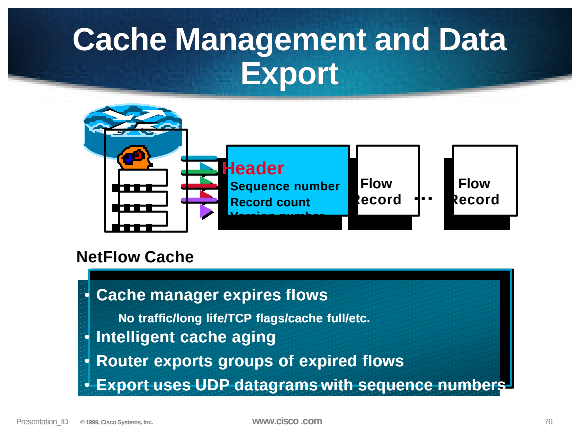


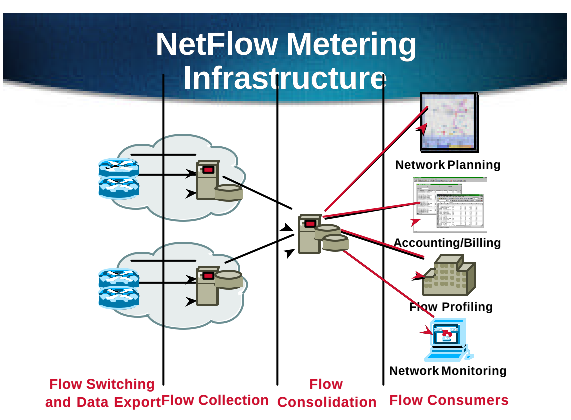


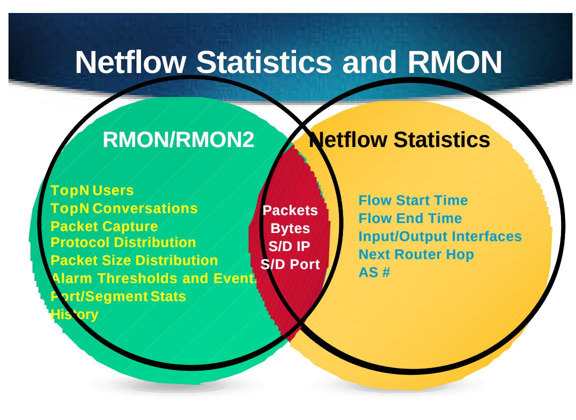


Protocol		Total Flows	Flows / Sec.	Packets / Flow	Bytes / Pkt	Packets / Sec.	Active Sec / Flow	Idle Sec / Flow
TCP -	Telnet	267,034	1.8	233	75	439.3	182.6	36.5
	FTP	1,030,837	7.2	10	78	76.6	22.6	43.7
	FTPD	554,967	3.9	164	345	641.3	52.7	15.7
	WWW	32,107,858	226.2	15	247	3610.6	13.5	28.1
	SMTP	3,526,231	24.8	13	159	323.1	10.2	23.6
	X	9,600	0.0	121	129	8.2	148.2	55.1
	BGP	111,096	0.7	14	77	11.5	229.2	61.1
	other	5,729,172	40.3	70	220	2858.1	71.0	41.3
UDP -	TFTP	2,398	0.0	3	62	0.0	13.4	69.5
	DNS	12,875,077	90.7	2	110	195.4	5.4	43.6
	other	1,489,072	10.4	30	293	321.8	28.5	68.7
ICMP		665,771	4.6	13	259	62.8	75.7	66.8
IGMP		5,144	0.0	18	278	0.6	82.4	64.3
IPINIP		4,450	0.0	933	377	29.2	166.7	61.0
IP -	other	2,693	0.0	11	136	0.2	80.8	65.7
TOTAL		58,381,400	411.3	20	227	8579.4	0.0	0.0

Src Intf	Source IP Address	Dest Intf	Dest IP Address	Prt	Src Port	Dest Port	Pkts	Bytes/ / Pkt	Act Sec / Flow	Idle Sec / Flow
Hs3/0	204.119.134.49	Fd0/0	142.35.4.36	6	0050	0610	1	44	0.0	0.6
Fd0/0	206.42.156.2	Hs3/0	206.52.126.29	6	0439	0050	12	105	9.1	1.0
Hs3/0	125.160.1.24	Fd0/0	200.246.225.8	6	BB81	0DB7	745	542	323.0	0.0
...	...	...	...	...	...	...	...	...	...	...

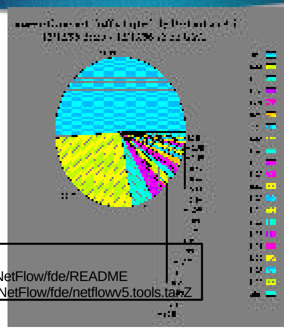






NetFlow Provides Open Interfaces

Cflowd
by ANS & BBN
and maintained by
CAIDA

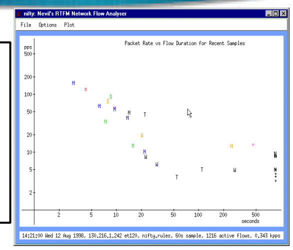


Http://www.caida.org
ftp://ftp-eng.cisco.com/ftp/NetFlow/fde/README
ftp://ftp-eng.cisco.com:/ftp/NetFlow/fde/netflowv5.tools.tar.Z

NetFlow Provides Open Interfaces

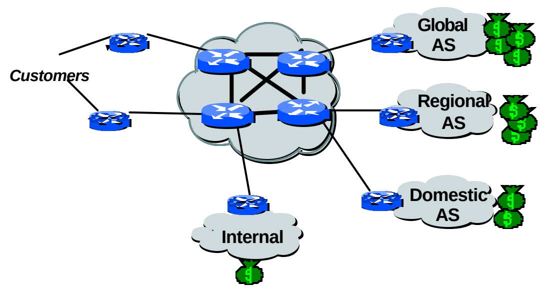
NetFlowMet
by Nevil Brownlee

- **Uses the work from the IETF's Realtime Traffic Flow Measurement (RTFM) WG**



http://www.auckland.ac.nz/net/Accounting/
http://www.auckland.ac.nz/net/Internet/rtfm/

NetFlow Distance-Based Accounting



NetFlow Distance-Based Accounting

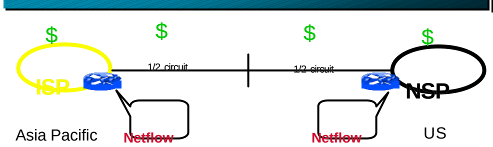
- **General Information page for Cisco Netflow services:**
  - ✓ http://www.cisco.com/warp/public/732/netflow/
- **Cisco's NetFlow FlowCollector v2.0 and NetFlow FlowAnalyzer v2.0:**
  - ✓ http://www.cisco.com/warp/public/732/netflow/netan_ov.htm
- **Case Study - How to implement Netflow in a network. *Traffic Accounting Using Netflow and Cflowd* by Roberto Sabatino (DANTE/TEN-34)**
  - ✓ http://www.dante.net/pubs/dip/32/32.html

NetFlow Distance-Based Accounting

- **3rd Party Solutions:**
  - ✓ Belle Systems http://www.belle.dk
  - ✓ Solect http://www.solect.com
  - ✓ XACCT Technologies http://www.xacct.com
  - ✓ Apogee Networks, Inc. http://www.Apogeenet.com
  - ✓ RODOPI http://www.rodopi.com

Netflow as a tool

Netflow statistics empowers all ISPs with the ability to know the who, what, where, and how much.



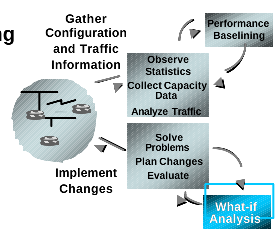
Presentation_ID.scr

Conclusions

- Aggressive measurement and analysis is critical to an ISPs and the Internet's survival.
- Not enough measurement and actual data analysis is taking place on the Internet. Too many people are speculating with weak data to back up their claims.

What-If Analysis

- Simulation applications
- Lab application modeling
  - ✓Protocol analyzer, WAN emulator, packet generator, NETSYS performance analyzer
- Lab network modeling
  - ✓NVS/NVT, lab network modeling



Service Level Management

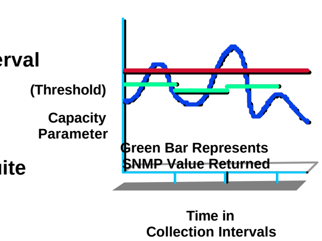
- Define performance requirements
- Define Upgrade criteria by capacity area
- Measure capacity and performance
- Review thresholds and baseline
- Take action!

Service Level Management

Threshold	WAN	LAN
CPU	75-90%	75-90%
Link	80-90%	40-90%
Memory	50%	50%
Output Queue	200	25
Buffer Misses	Any	Any
Broadcast Vol	10/Sec	300/Sec
FECN/BECN	10/Sec	N/A

Peak and Average Utilization

- Solution to narrow collection interval
- Low collection interval = high overhead
- Recommend >=5 minutes
- Peak values not quite what they seem
- Close to threshold indicates likely exceed condition

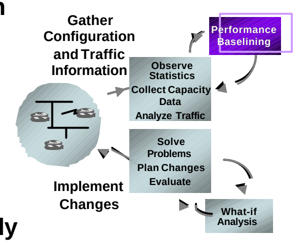


Capacity Exception Management

- Alarm critical capacity thresholds (CPU, critical link)
- Develop notification, escalation and action plan for threshold violations
- Take action!

Performance and Capacity Baselining

- Interface utilization
- Device CPU, memory, buffer, I/O utilization
- Network overhead
- Raw performance characteristics
- Monthly or quarterly baseline report



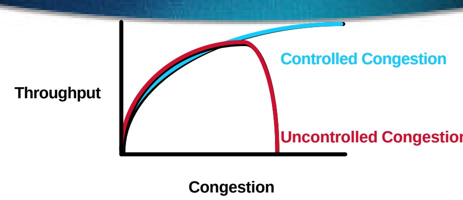
Upgrade Planning

- Understand lead times for circuits, equipment, planning and design
- upgrade criteria based on service level management

QoS Management

- Prioritize applications by business impact
- Understand networked application behavior (packet size, timeouts, flows, bandwidth requirements)
- Develop QoS management plan

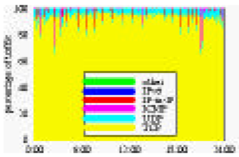
Problems with Congestion

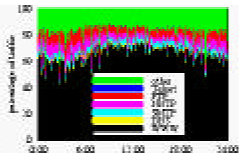


- Uncontrolled congestion will seriously degrade system performance
  - ✓ The system buffers fill up
  - ✓ Packets are dropped, resulting in retransmissions
  - ✓ This causes more packet loss and increased latency
  - ✓ The problem builds on itself until the system collapses

Backbone Traffic Mix

Transport Breakout | TCP Applications





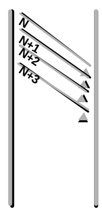
Source: MCI/NSF OC-3MON via http://www.nlanr.net, 1998

TCP Technology Issues

- Single drops communicate from network to sending host
  - ✓ "You need to slow down"
- Multiple drops in round trip trigger time-outs
  - ✓ "Something bad happened out here"

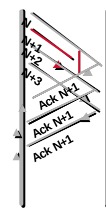
Behavior of a TCP Sender

- Sends as much as credit allows
- Starts credit small
  - ✓ Avoid overloading network queues
- Increases credit exponentially
  - ✓ To gauge network capability



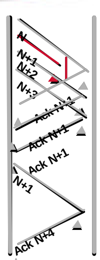
Behavior of a TCP Receiver

- When in receipt of "next message," schedules an ACK
- When in receipt of something else, acknowledges all it can immediately



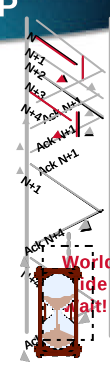
Sender Response to ACK

- If ACK acknowledges something
  - ✓ Update credit and send
- If not, presume it indicates a lost packet
  - ✓ Send first unacknowledged message right away
  - ✓ Halve current credit
  - ✓ Increase linearly to gauge network throughput



Multiple Drops in TCP

- In the event of multiple drops within the same session:
  - ✓ Current TCPs wait for time-out
  - ✓ Selective acknowledge may work around (but see INFOCOM '98)
  - ✓ New Reno "fast retransmit phase" takes several RTTs to recover



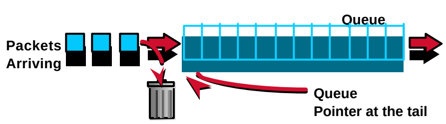
How Can We Make TCP in a Network Act Predictably?

- Predictable amount of traffic in the network:
  - ✓ Well-written TCP implementations manage their rates to the available bandwidth
- Router needs to
  - ✓ Provide predictable treatment of packets
  - ✓ Queue delay and drop characteristics

Fundamental FIFO Queue Management Technologies

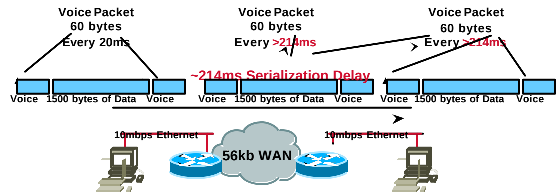
- Tail drop
  - ✓ Network standard behavior
  - ✓ Causes session synchronization when waves of traffic experience correlated drops
- Random Early Detection (RED)
  - ✓ Random drops used to desynchronize TCP sessions and control rates

Tail Drop



- Without RED, when the queue fills up all packets that arrive are dropped—Tail drop

Large Packets "Freeze Out" Voice

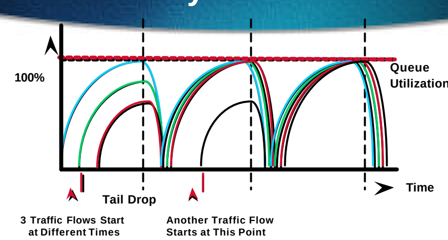


- Large packets can cause playback buffer underrun, tail drops
- Jitter or playback buffer can accommodate some delay/delay variation

Session Synchronization

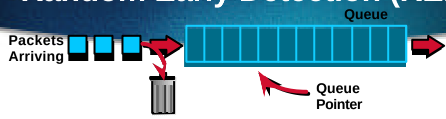
- Session synchronization results from synchronized losses
- Tail drop from waves of traffic synchronizes losses

Session Synchronization



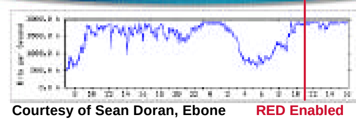
- TCP is a sliding window protocol that uses self-clocking to adjust its use of the network to match available bandwidth using slow-start and congestion avoidance algorithm
- Session synchronisation is when many TCP connections go through TCP Slow-Start mode at the same time

Random Early Detection (RED)



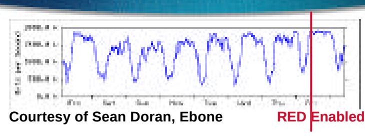
- With RED, as oppose to doing a tail drop,
  the router monitors the **average queue size** and using randomization it chooses connections to notify that a congestion is impending
  - ✓ Note: **Avg. queue size** is not an

Effect of Random Early Detection



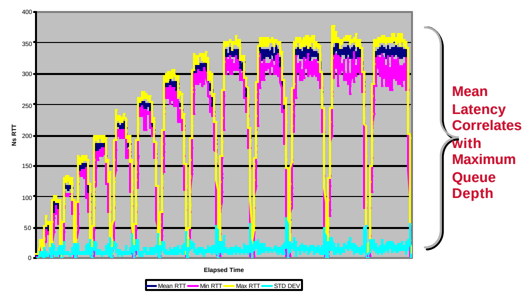
- One day, below 100% throughput
  - ✓ Simple FIFO with tail drop
- Starting 10:00 second day, 100% throughput
  - ✓ Random Early Detection enabled

Was that a Fluke?

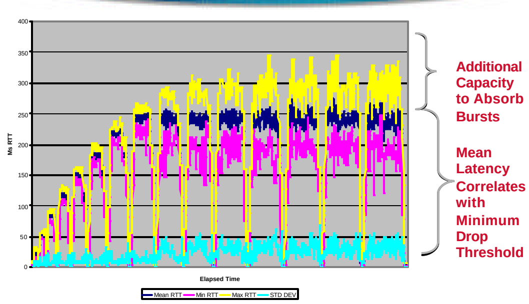


- No, here's what happened that week…
- Session synchronization reduced throughput until RED enabled

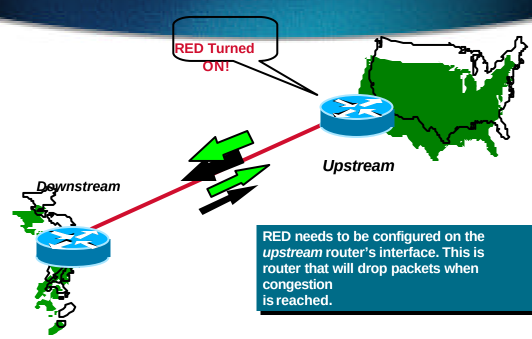
FIFO Traffic Timings



RED Traffic Timings



Where to apply RED?



Applying RED/WRED

- Enabling WRED
  - ✓ [no] random-detect <weight-constant>
  - ✓ weight-constant = <1-16> is an integer used in weighted average to mean 2^weigh-constant. 10 is the default.
- Tuning weight constant affects loss rate
  - ✓ rule-of-thumb:
  - ✓ DS-3/OC-3 Links: Value of 10 might achieve ~10^-4 drop rate, recommended for DS-3/OC-3 link.
  - ✓ T1/E1 Links: Value of 7 might achieve a loss rate around 10^-3.
  - ✓ Actual recommended value should be determined in real operational network.

Therefore—TCP QoS Definition:

- Normally at most one drop per round trip
- Mean variation in latency bounded by predictable network

TCP Flow Statistics

- >90% of sessions have ten packets each way or less
  - ✓ Transaction mode (mail, small web page)
- >80% of all TCP traffic results from <10% of the sessions, in high rate bursts
  - ✓ It is these that we worry about managing

An Interesting Common Fallacy about RED:

- "RED means you will have more drops"
  - ✓ Statement derives from observed statistics
- RED means that you will have
  - ✓ Closer to 100% utilization of your line
  - ✓ Less average delay per packet
- But queuing theory?
  - ✓ As a line approaches 100% utilization, drops will increase, even though served load increases

TCP Traffic Management Issues

- Applications
  - ✓ Often have site-specific policy associated with them
  - ✓ Traffic often identifiable by port numbers
- Sites
  - ✓ Generally identifiable by address prefix or interface traffic is received on

TCP Bandwidth Policy Questions to Answer

- Particular site or application wants at least a certain bandwidth
- Particular site or application wants at most a certain bandwidth
- Particular site or application wants to average about a certain bandwidth

What to look for in an Upstream Provider

Preparation

- List all potential providers
- Get Maps of oceanic cable systems
- Get Maps of satellite foot prints.

What to ask from the prospective providers?

- Network Maps with landing/termination points of your links.
- List of IXPs and Private Peers
- URLs of NOC Pages
- Do they lease routers and/or co-locations space?
- Do they have upstream caches?

Example - Network Map

What to require from the Upstream Provider

- Statistics Page and Weekly Reports
- 24x7 NOC Contacts
- RED or WRED on their router's interface
- CAR ICMP Rates Limits for DoS Protection
- Back-up contingencies in writing.

CISCO SYSTEMS

EMPOWERING THE INTERNET GENERATION℠