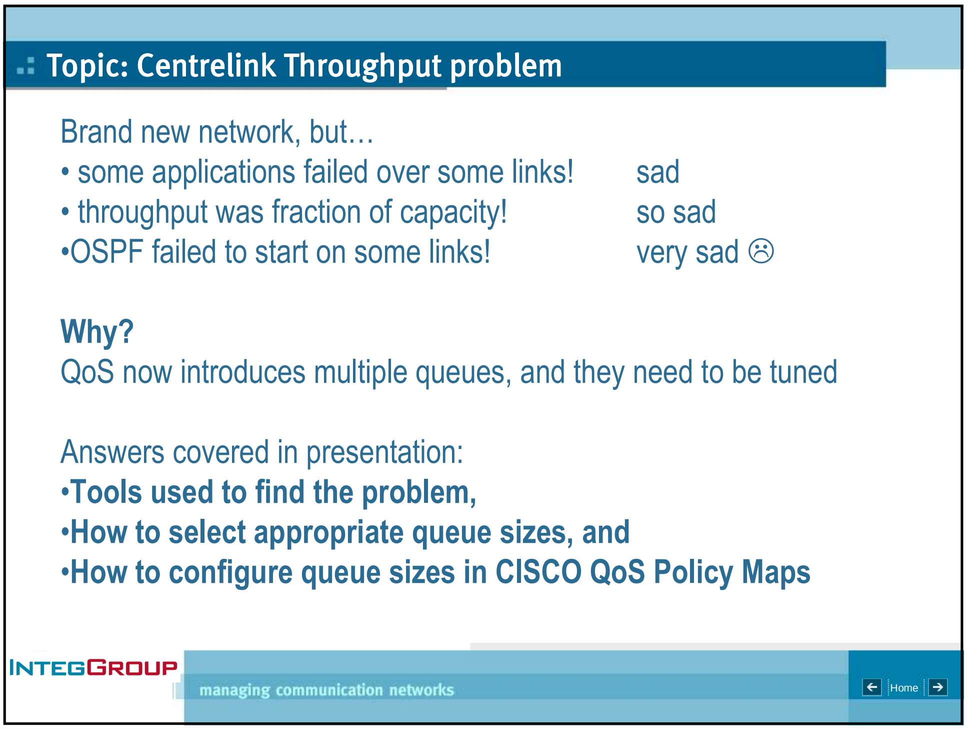
Case Study: Centrelink queue length

Stephen Kingham
QESTNET 2008 (www.qestnet.net.au)
Royal Pines Queensland Australia

INTEGGROUP

managing communication networks



Topic: Centrelink Throughput problem

Brand new network, but…

• some applications failed over some links!	sad
• throughput was fraction of capacity!	so sad
•OSPF failed to start on some links!	very sad ☹

Why?

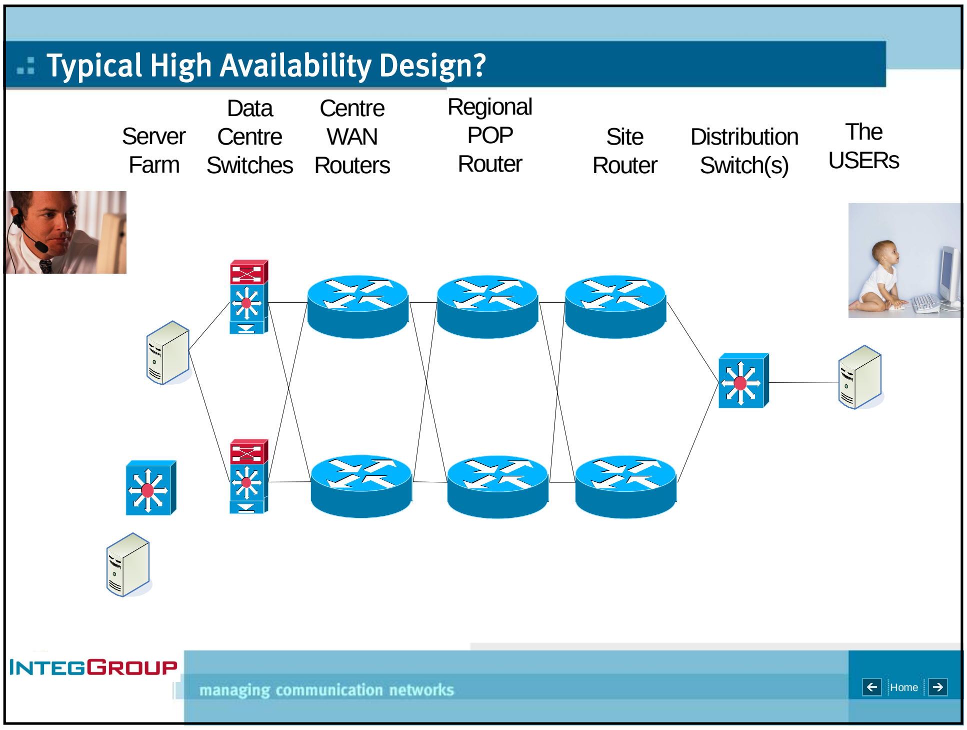
QoS now introduces multiple queues, and they need to be tuned

Answers covered in presentation:

•Tools used to find the problem,
•How to select appropriate queue sizes, and
•How to configure queue sizes in CISCO QoS Policy Maps

INTEGGROUP

managing communication networks



Typical High Availability Design?

Server Farm

Data Centre Switches

Centre WAN Routers

Regional POP Router

Site Router

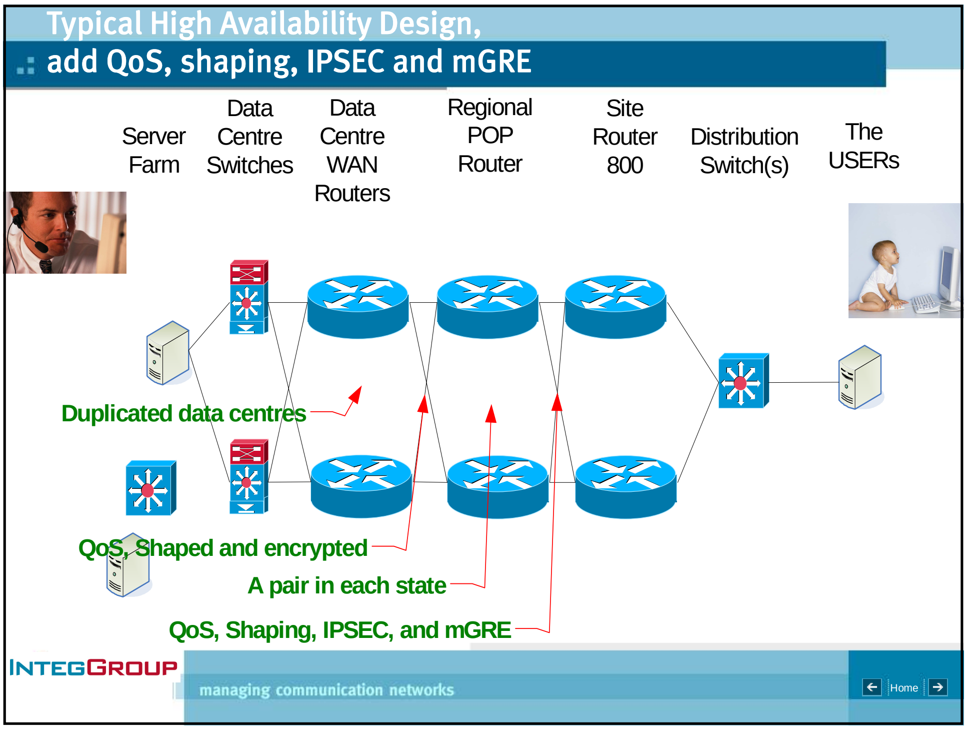
Distribution Switch(s)

The USERs

INTEGGROUP

managing communication networks

Home



Typical High Availability Design, add QoS, shaping, IPSEC and mGRE

Server Farm

Data Centre Switches

Data Centre WAN Routers

Regional POP Router

Site Router 800

Distribution Switch(s)

The USERs

Duplicated data centres

QoS, Shaped and encrypted

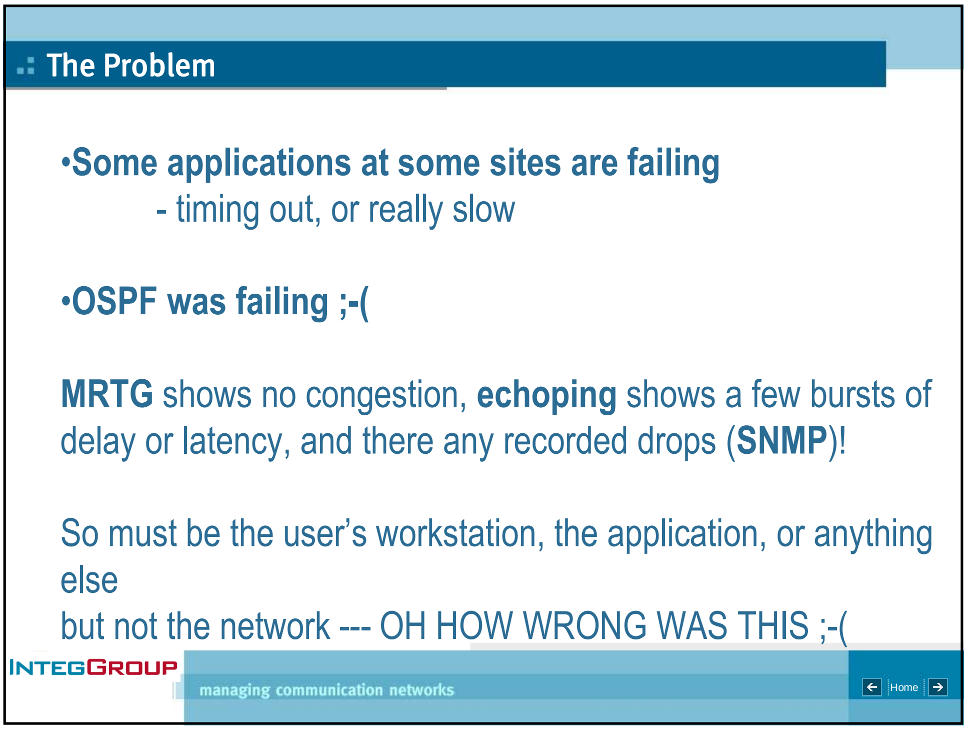
A pair in each state

QoS, Shaping, IPSEC, and mGRE

INTEGGROUP

managing communication networks

Home



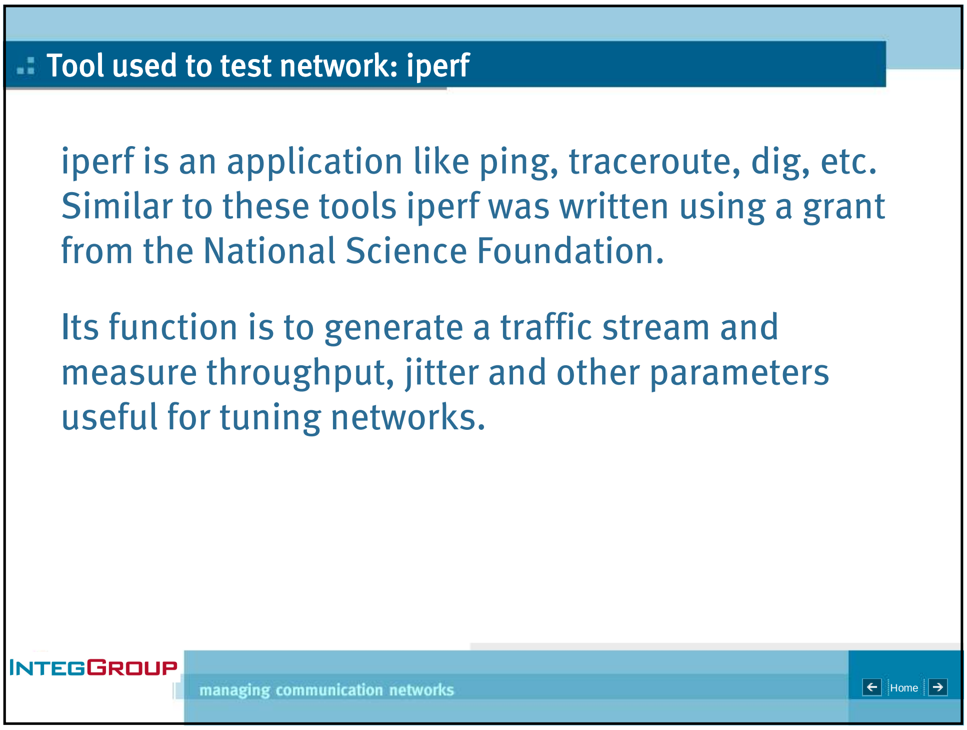
The Problem

- **Some applications at some sites are failing**
  - timing out, or really slow

- **OSPF was failing ;-(**

MRTG shows no congestion, **echoping** shows a few bursts of delay or latency, and there any recorded drops (**SNMP**)!

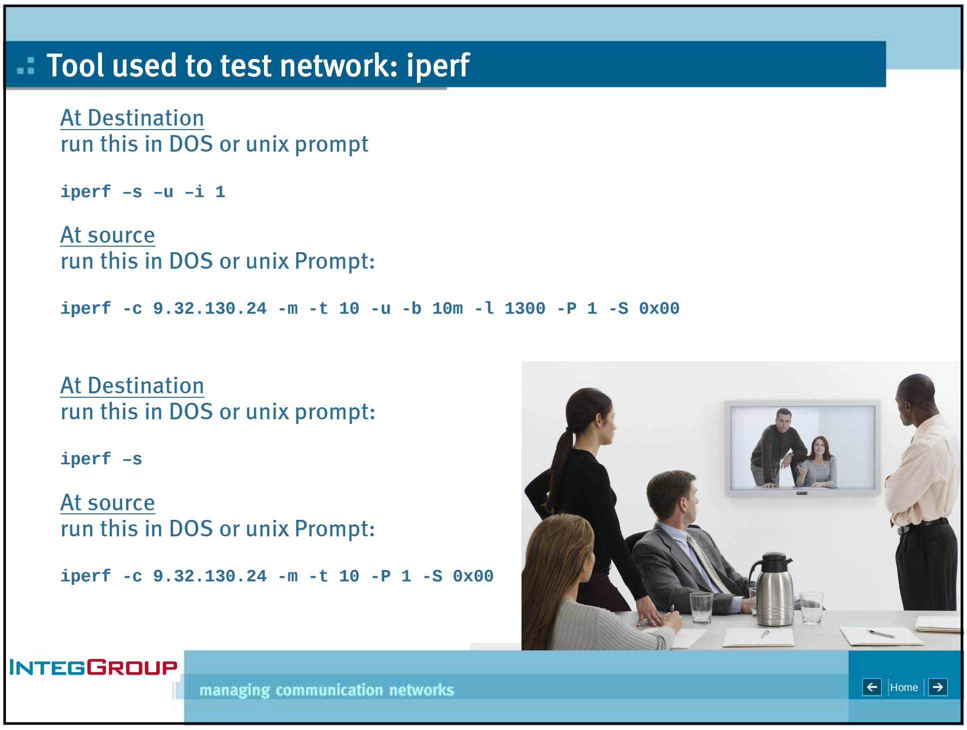
So must be the user’s workstation, the application, or anything else
but not the network --- OH HOW WRONG WAS THIS ;-(



Tool used to test network: iperf

iperf is an application like ping, traceroute, dig, etc. Similar to these tools iperf was written using a grant from the National Science Foundation.

Its function is to generate a traffic stream and measure throughput, jitter and other parameters useful for tuning networks.



Tool used to test network: iperf

At Destination
run this in DOS or unix prompt

```
iperf –s –u –i 1
```

At source
run this in DOS or unix Prompt:

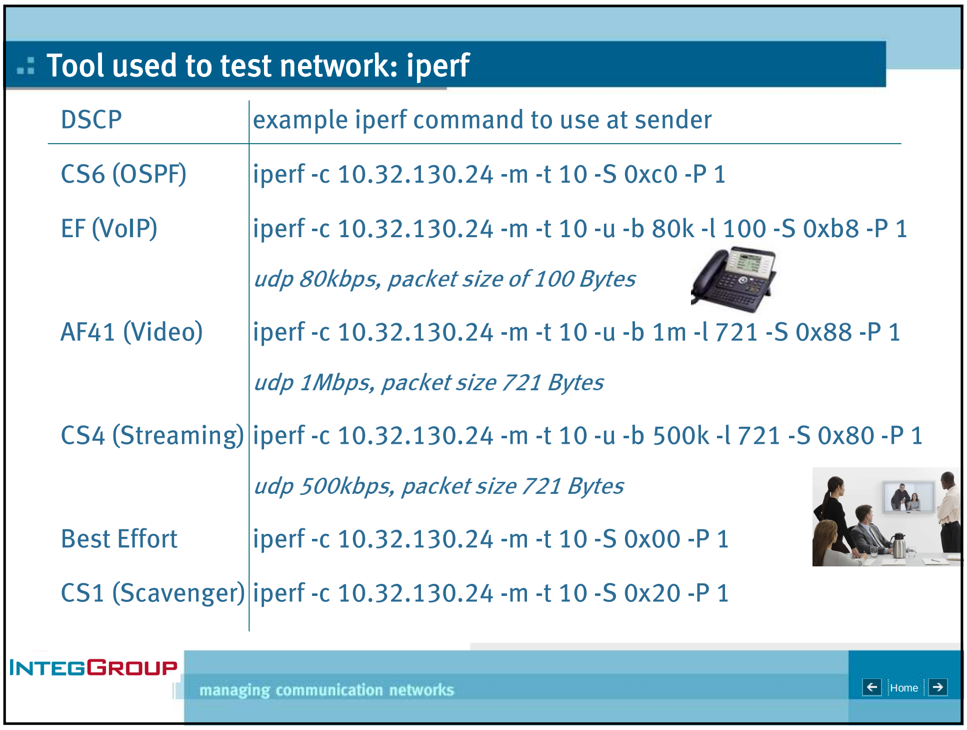
```
iperf -c 9.32.130.24 -m -t 10 -u -b 10m -l 1300 -P 1 -S 0x00
```

At Destination
run this in DOS or unix prompt:

```
iperf –s
```

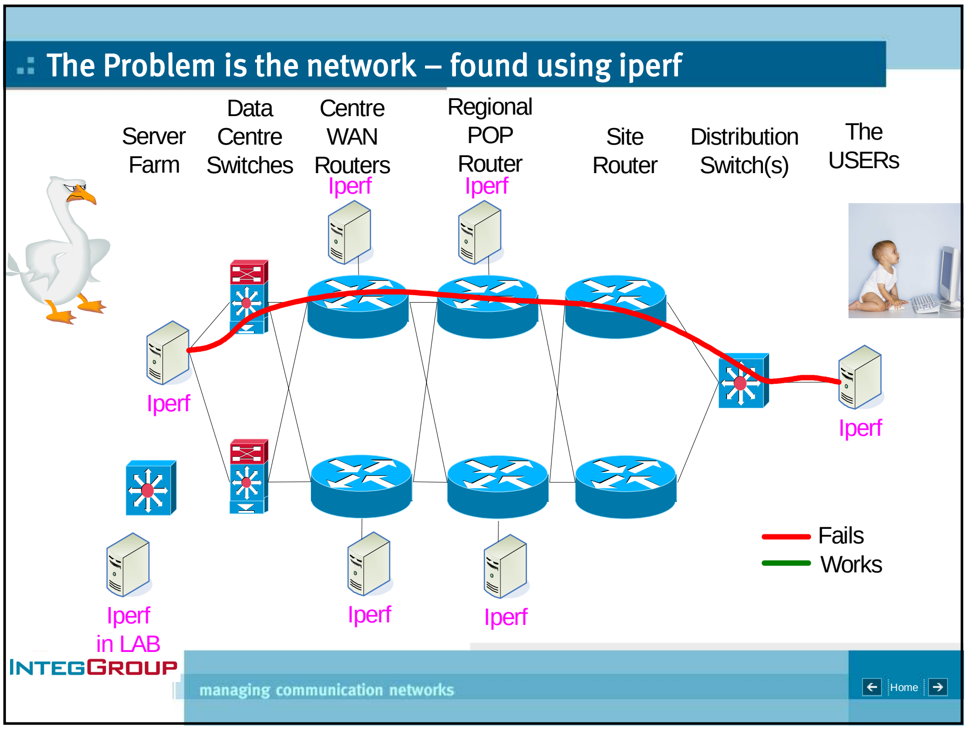
At source
run this in DOS or unix Prompt:

```
iperf -c 9.32.130.24 -m -t 10 -P 1 -S 0x00
```



Tool used to test network: iperf

DSCP	example iperf command to use at sender
CS6 (OSPF)	iperf -c 10.32.130.24 -m -t 10 -S 0xc0 -P 1
EF (VoIP)	iperf -c 10.32.130.24 -m -t 10 -u -b 80k -l 100 -S 0xb8 -P 1
	udp 80kbps, packet size of 100 Bytes
AF41 (Video)	iperf -c 10.32.130.24 -m -t 10 -u -b 1m -l 721 -S 0x88 -P 1
	udp 1Mbps, packet size 721 Bytes
CS4 (Streaming)	iperf -c 10.32.130.24 -m -t 10 -u -b 500k -l 721 -S 0x80 -P 1
	udp 500kbps, packet size 721 Bytes
Best Effort	iperf -c 10.32.130.24 -m -t 10 -S 0x00 -P 1
CS1 (Scavenger)	iperf -c 10.32.130.24 -m -t 10 -S 0x20 -P 1



The Problem is the network – found using iperf

Server Farm

Data Centre Switches

Centre WAN Routers
Iperf

Regional POP Router
Iperf

Site Router

Distribution Switch(s)

The USERs

Iperf

Iperf

Iperf

Iperf

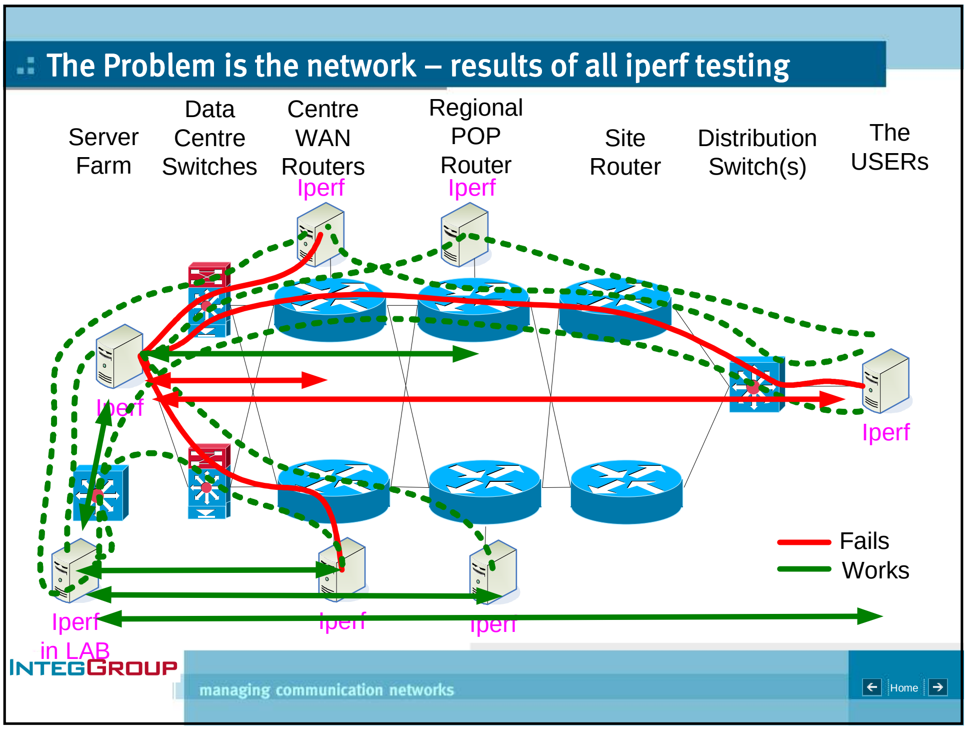
Iperf
in LAB

🟥 Fails
🟩 Works

INTEGGROUP

managing communication networks

← Home →



The Problem is the network – results of all iperf testing

Server Farm

Data Centre Switches

Centre WAN Routers
Iperf

Regional POP Router
Iperf

Site Router

Distribution Switch(s)

The USERs

Iperf

Iperf

Iperf

Iperf

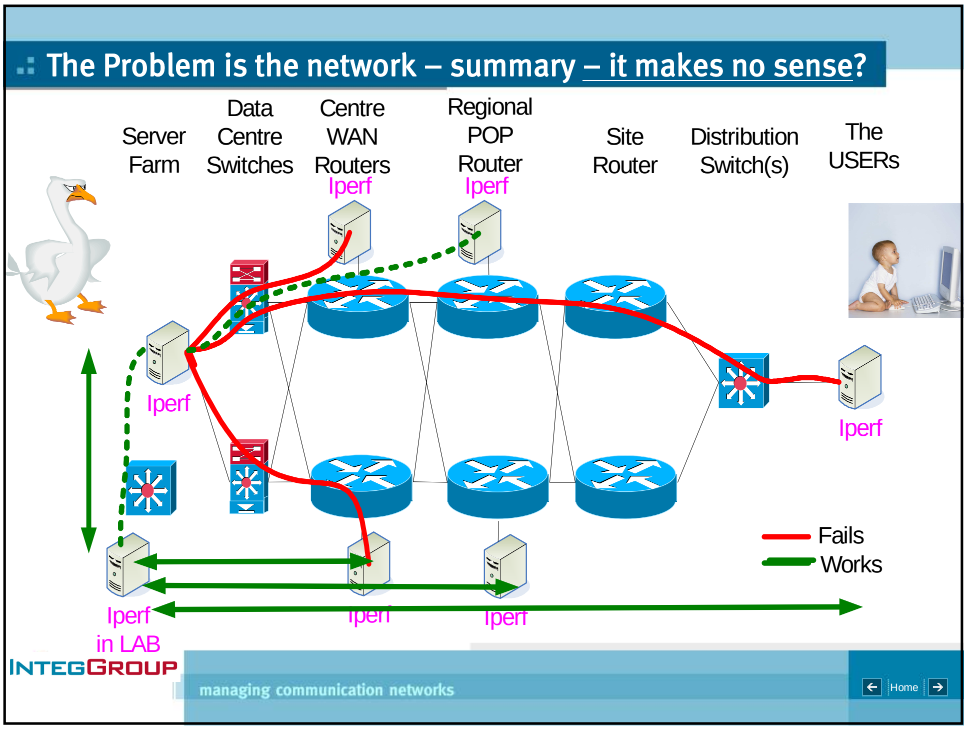
Iperf
in LAB

🟥 Fails
🟩 Works

INTEGGROUP

managing communication networks

← Home →

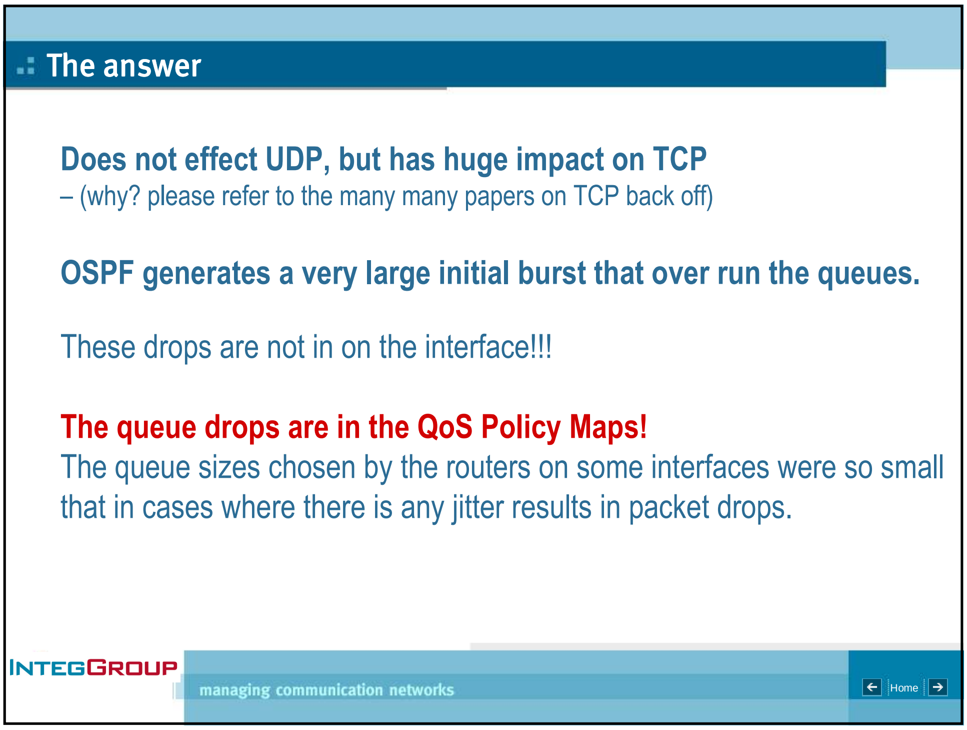


The Problem is the network – summary – it makes no sense?

INTEGGROUP

managing communication networks

Home



The answer

Does not effect UDP, but has huge impact on TCP

– (why? please refer to the many many papers on TCP back off)

OSPF generates a very large initial burst that over run the queues.

These drops are not in on the interface!!!

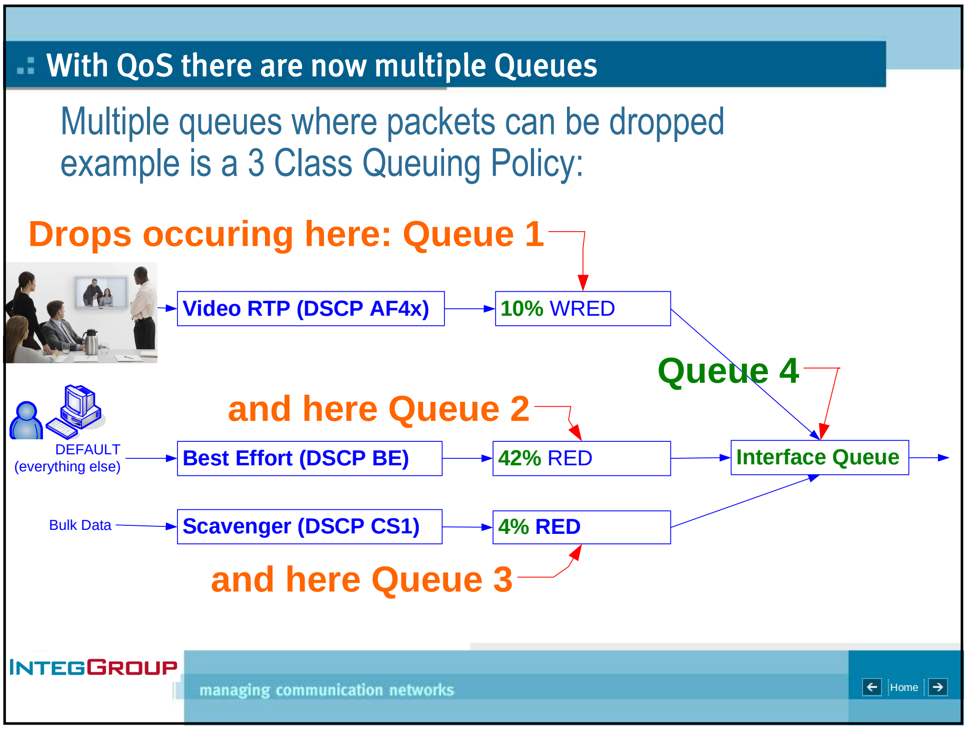
The queue drops are in the QoS Policy Maps!

The queue sizes chosen by the routers on some interfaces were so small that in cases where there is any jitter results in packet drops.

INTEGGROUP

managing communication networks

Home



With QoS there are now multiple Queues

Multiple queues where packets can be dropped
example is a 3 Class Queuing Policy:

Drops occuring here: Queue 1

Video RTP (DSCP AF4x) → 10% WRED

Queue 4

and here Queue 2

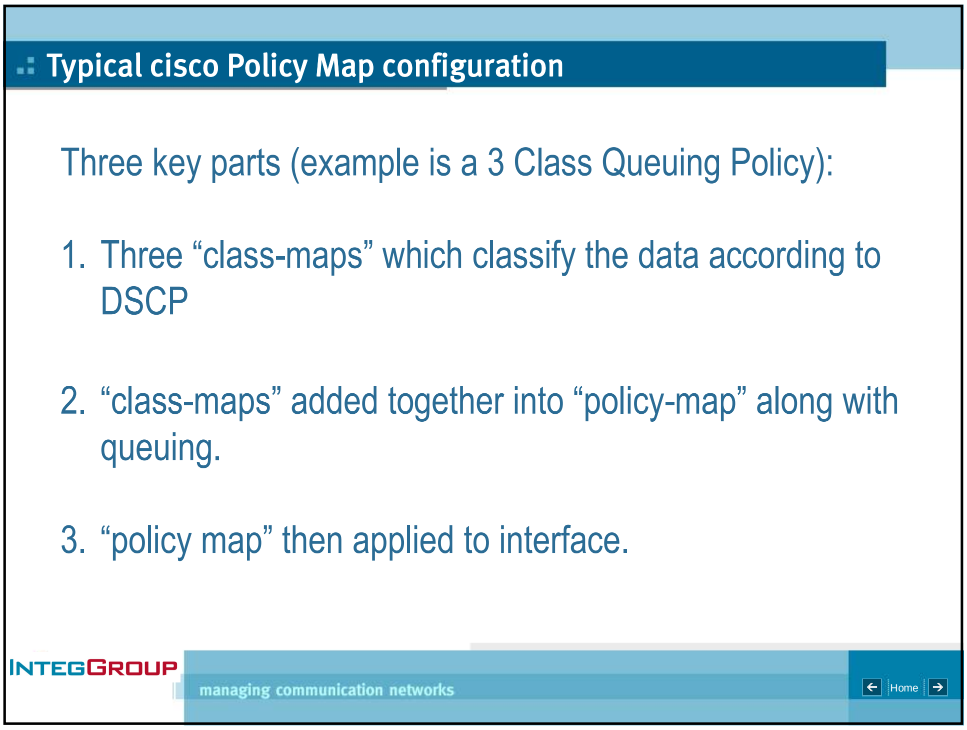
DEFAULT
(everything else) → Best Effort (DSCP BE) → 42% RED → Interface Queue →

Bulk Data → Scavenger (DSCP CS1) → 4% RED

and here Queue 3

INTEGGROUP managing communication networks

Home



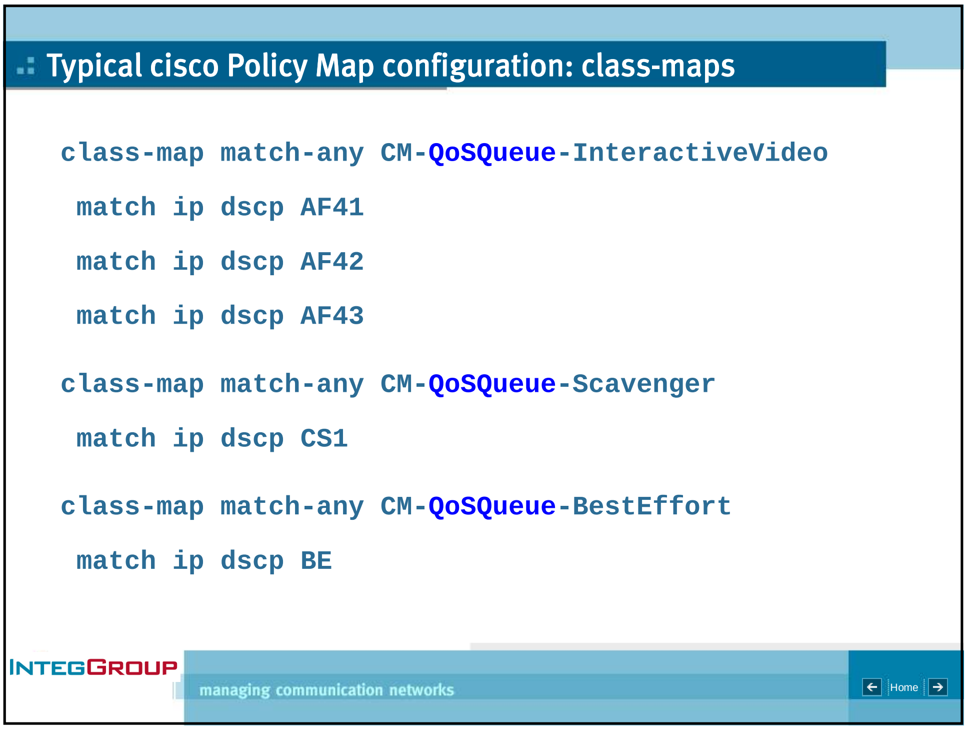
Typical cisco Policy Map configuration

Three key parts (example is a 3 Class Queuing Policy):

1. Three “class-maps” which classify the data according to DSCP
2. “class-maps” added together into “policy-map” along with queuing.
3. “policy map” then applied to interface.

INTEGGROUP managing communication networks

Home

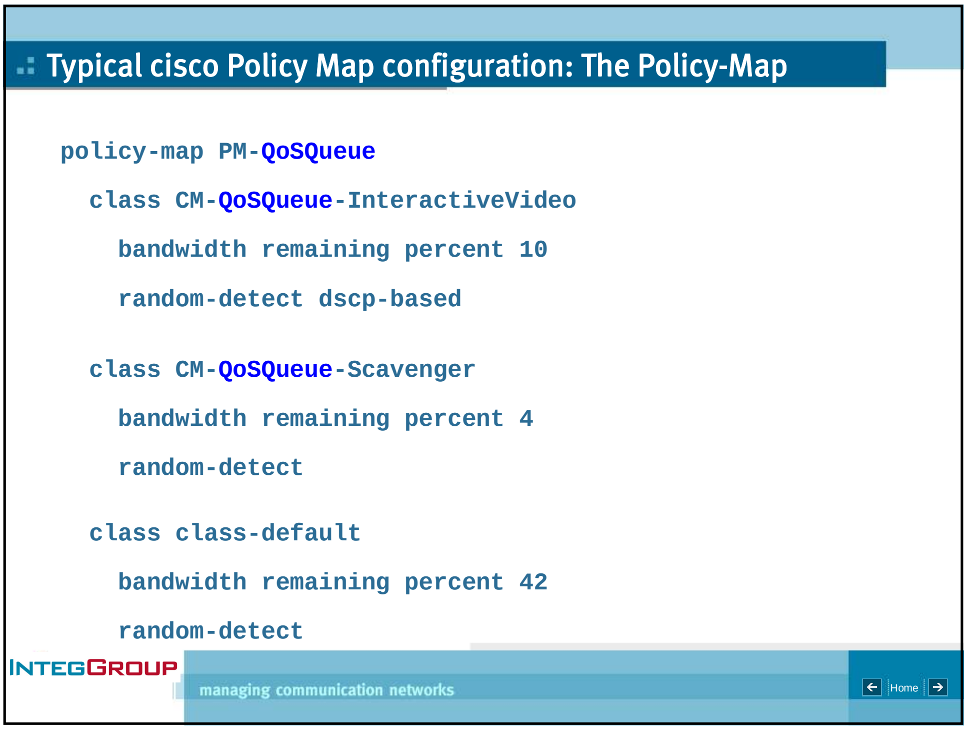


Typical cisco Policy Map configuration: class-maps

```
class-map match-any CM-QoSQueue-InteractiveVideo
 match ip dscp AF41
 match ip dscp AF42
 match ip dscp AF43
class-map match-any CM-QoSQueue-Scavenger
 match ip dscp CS1
class-map match-any CM-QoSQueue-BestEffort
 match ip dscp BE
```

INTEGGROUP managing communication networks

Home

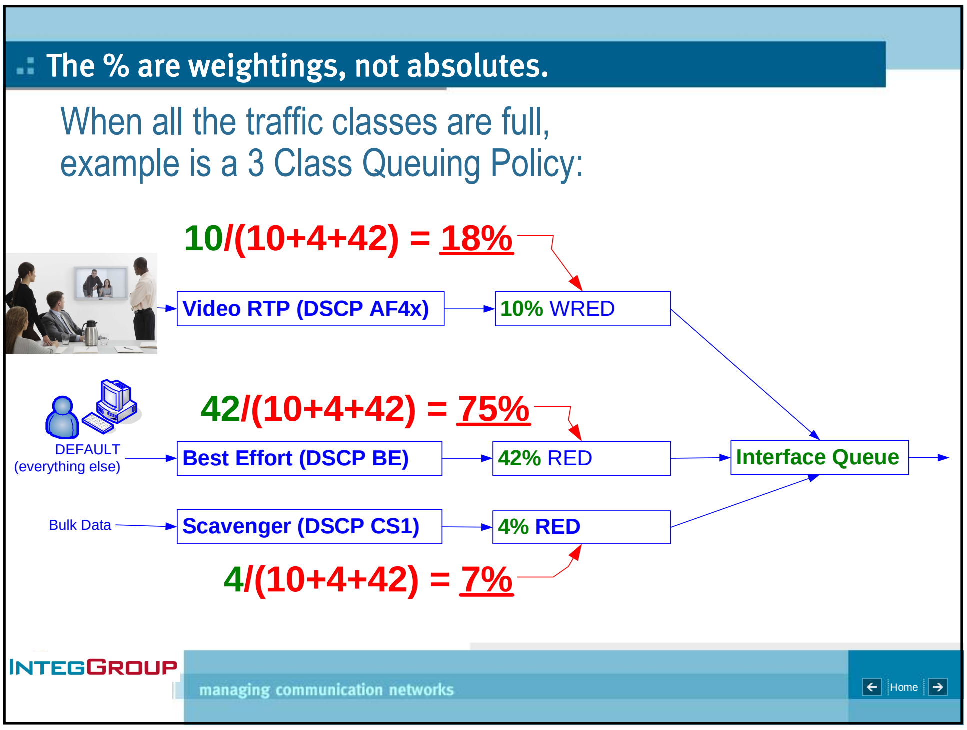


Typical cisco Policy Map configuration: The Policy-Map

```
policy-map PM-QoSQueue
  class CM-QoSQueue-InteractiveVideo
    bandwidth remaining percent 10
    random-detect dscp-based
  class CM-QoSQueue-Scavenger
    bandwidth remaining percent 4
    random-detect
  class class-default
    bandwidth remaining percent 42
    random-detect
```

INTEGGROUP managing communication networks

Home



The % are weightings, not absolutes.

When all the traffic classes are full,
example is a 3 Class Queuing Policy:

10/(10+4+42) = 18%

Video RTP (DSCP AF4x) → 10% WRED

DEFAULT (everything else) → Best Effort (DSCP BE) → 42% RED → Interface Queue →

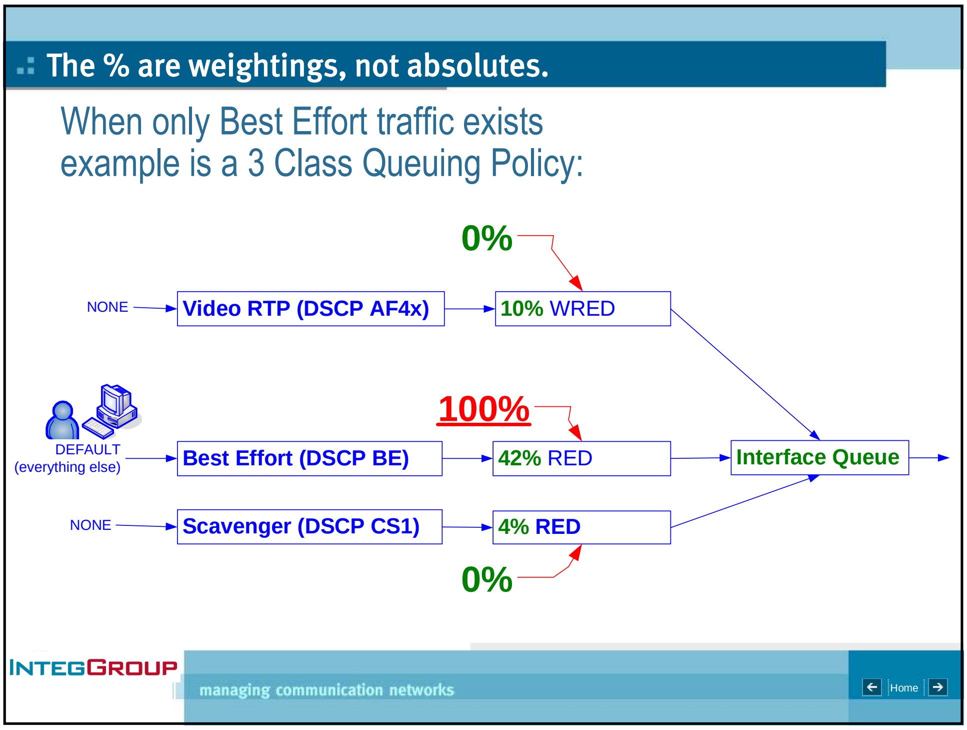
42/(10+4+42) = 75%

Bulk Data → Scavenger (DSCP CS1) → 4% RED

4/(10+4+42) = 7%

INTEGROUP

managing communication networks



The % are weightings, not absolutes.

When only Best Effort traffic exists
example is a 3 Class Queuing Policy:

0%

NONE → Video RTP (DSCP AF4x) → 10% WRED

100%

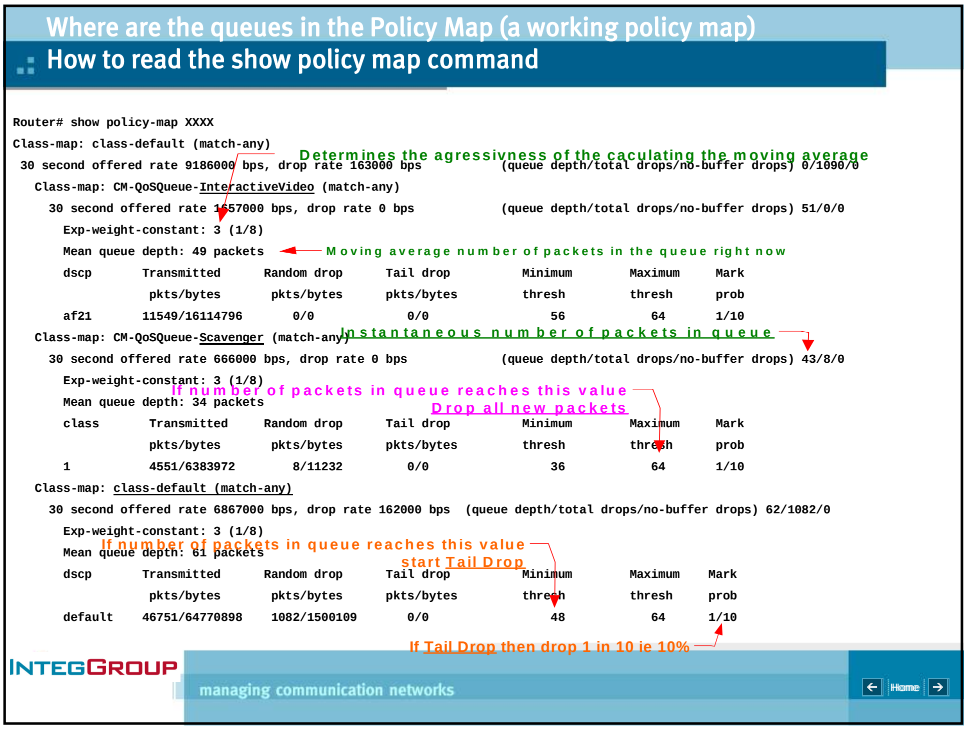
DEFAULT (everything else) → Best Effort (DSCP BE) → 42% RED → Interface Queue →

NONE → Scavenger (DSCP CS1) → 4% RED

0%

INTEGROUP

managing communication networks



Where are the queues in the Policy Map (a working policy map)

How to read the show policy map command

```
Router# show policy-map XXXX
Class-map: class-default (match-any)
 30 second offered rate 9186000 bps, drop rate 163000 bps          (queue depth/total drops/no-buffer drops) 0/1090/0
   Class-map: CM-QoSQueue-InteractiveVideo (match-any)
     30 second offered rate 1657000 bps, drop rate 0 bps            (queue depth/total drops/no-buffer drops) 51/0/0
       Exp-weight-constant: 3 (1/8)
       Mean queue depth: 49 packets
       dscp       Transmitted      Random drop      Tail drop          Minimum        Maximum     Mark
                   pkts/bytes       pkts/bytes      pkts/bytes          thresh         thresh      prob
       af21       11549/16114796       0/0             0/0                56             64        1/10
   Class-map: CM-QoSQueue-Scavenger (match-any)
     30 second offered rate 666000 bps, drop rate 0 bps             (queue depth/total drops/no-buffer drops) 43/8/0
       Exp-weight-constant: 3 (1/8)
       Mean queue depth: 34 packets
       class       Transmitted     Random drop      Tail drop          Minimum        Maximum     Mark
                   pkts/bytes       pkts/bytes      pkts/bytes          thresh         thresh      prob
       1           4551/6383972       8/11232          0/0                36             64        1/10
   Class-map: class-default (match-any)
     30 second offered rate 6867000 bps, drop rate 162000 bps  (queue depth/total drops/no-buffer drops) 62/1082/0
       Exp-weight-constant: 3 (1/8)
       Mean queue depth: 61 packets
       dscp       Transmitted      Random drop      Tail drop          Minimum        Maximum    Mark
                   pkts/bytes       pkts/bytes      pkts/bytes          thresh         thresh     prob
       default    46751/64770898    1082/1500109       0/0                48             64       1/10
```

Determines the agressivness of the caculating the moving average

Moving average number of packets in the queue right now

Instantaneous number of packets in queue

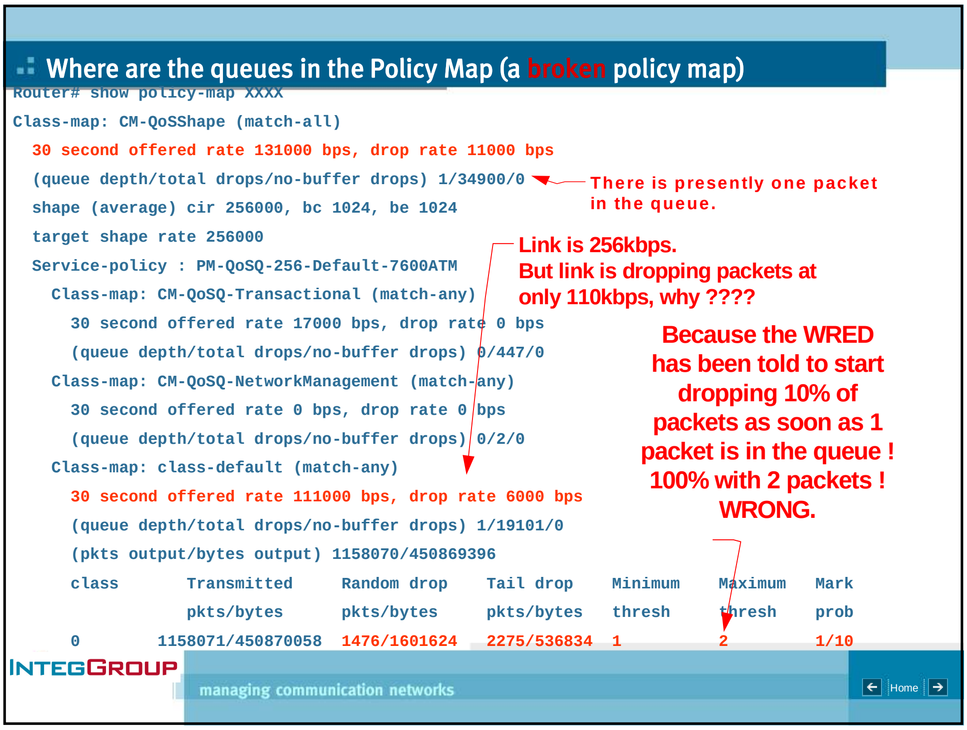
If number of packets in queue reaches this value Drop all new packets

If number of packets in queue reaches this value start Tail Drop

If Tail Drop then drop 1 in 10 ie 10%

INTEGGROUP

managing communication networks



Where are the queues in the Policy Map (a broken policy map)

```
Router# show policy-map XXXX
Class-map: CM-QoSShape (match-all)
  30 second offered rate 131000 bps, drop rate 11000 bps
  (queue depth/total drops/no-buffer drops) 1/34900/0
  shape (average) cir 256000, bc 1024, be 1024
  target shape rate 256000
  Service-policy : PM-QoSQ-256-Default-7600ATM
    Class-map: CM-QoSQ-Transactional (match-any)
      30 second offered rate 17000 bps, drop rate 0 bps
      (queue depth/total drops/no-buffer drops) 0/447/0
    Class-map: CM-QoSQ-NetworkManagement (match-any)
      30 second offered rate 0 bps, drop rate 0 bps
      (queue depth/total drops/no-buffer drops) 0/2/0
    Class-map: class-default (match-any)
      30 second offered rate 111000 bps, drop rate 6000 bps
      (queue depth/total drops/no-buffer drops) 1/19101/0
      (pkts output/bytes output) 1158070/450869396
      class       Transmitted       Random drop    Tail drop     Minimum     Maximum    Mark
                  pkts/bytes        pkts/bytes     pkts/bytes    thresh      thresh     prob
      0        1158071/450870058   1476/1601624   2275/536834   1           2          1/10
```

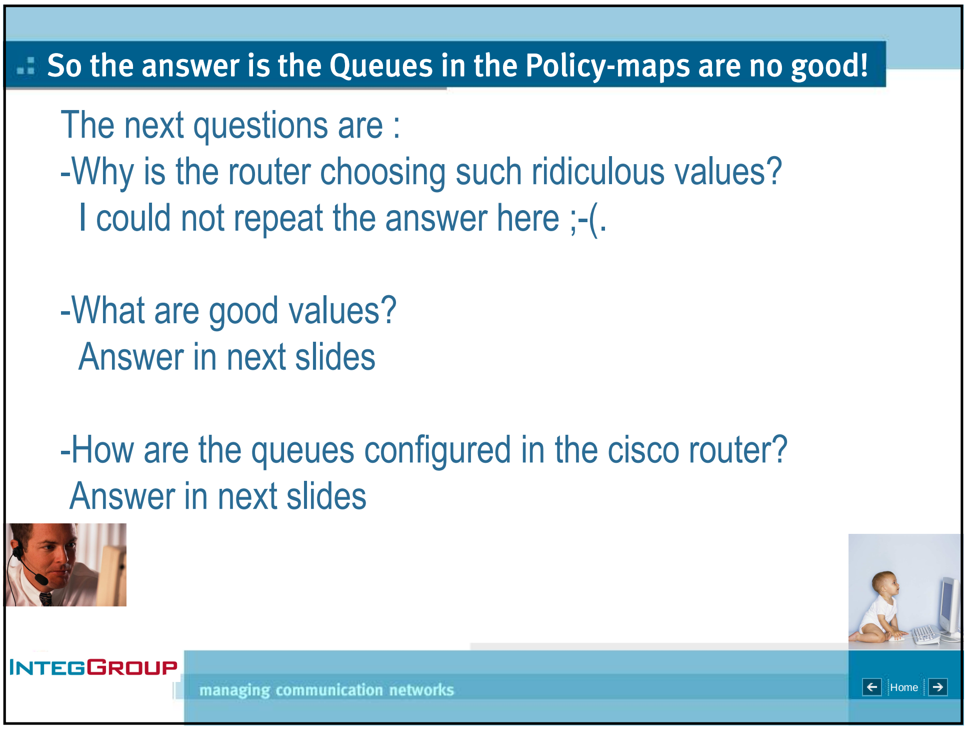
There is presently one packet in the queue.

Link is 256kbps.
But link is dropping packets at
only 110kbps, why ????

Because the WRED has been told to start dropping 10% of packets as soon as 1 packet is in the queue ! 100% with 2 packets ! WRONG.

INTEGGROUP

managing communication networks



So the answer is the Queues in the Policy-maps are no good!

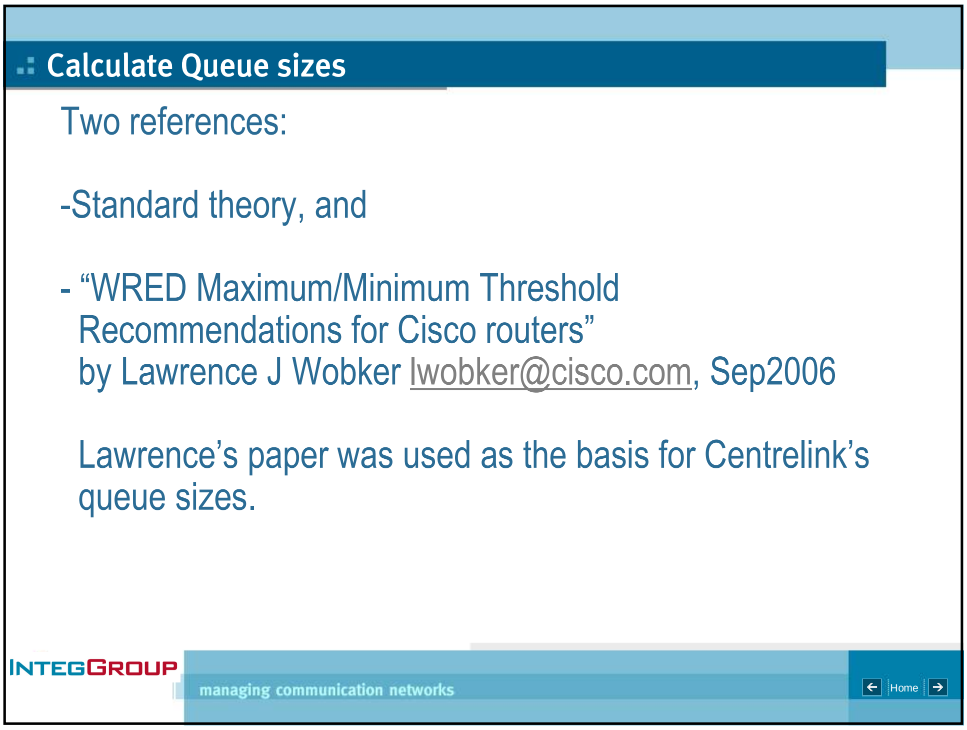
The next questions are :

-Why is the router choosing such ridiculous values?
I could not repeat the answer here ;-(.

-What are good values?
Answer in next slides

-How are the queues configured in the cisco router?
Answer in next slides

INTEGGROUP
managing communication networks
Home



Calculate Queue sizes

Two references:

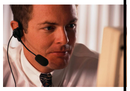
-Standard theory, and

- “WRED Maximum/Minimum Threshold Recommendations for Cisco routers”
by Lawrence J Wobker lwobker@cisco.com, Sep2006

Lawrence’s paper was used as the basis for Centrelink’s queue sizes.

INTEGGROUP
managing communication networks
Home

Calculate Queue sizes: Simple theory



Calculating the right size of a queue is based on

1. Bandwidth of link

2. Average packet size
   (eg Bulk 1400Bytes, InteractiveVideo is 721 Bytes)

3. Maximum delay through a congested queue
   (eg Bulk 1,000msec, InteractiveVideo 20msec)

Using simple theory:

Queue size = $(\text{Delay}_{\text{sec}} * \text{BandWidth}_{\text{bits/sec}}) / (\text{PacketSize}_{\text{Bytes}} * 8_{\text{bits}})$

Queue size for Interactive Video would be 34 packets on a 10Mbps link

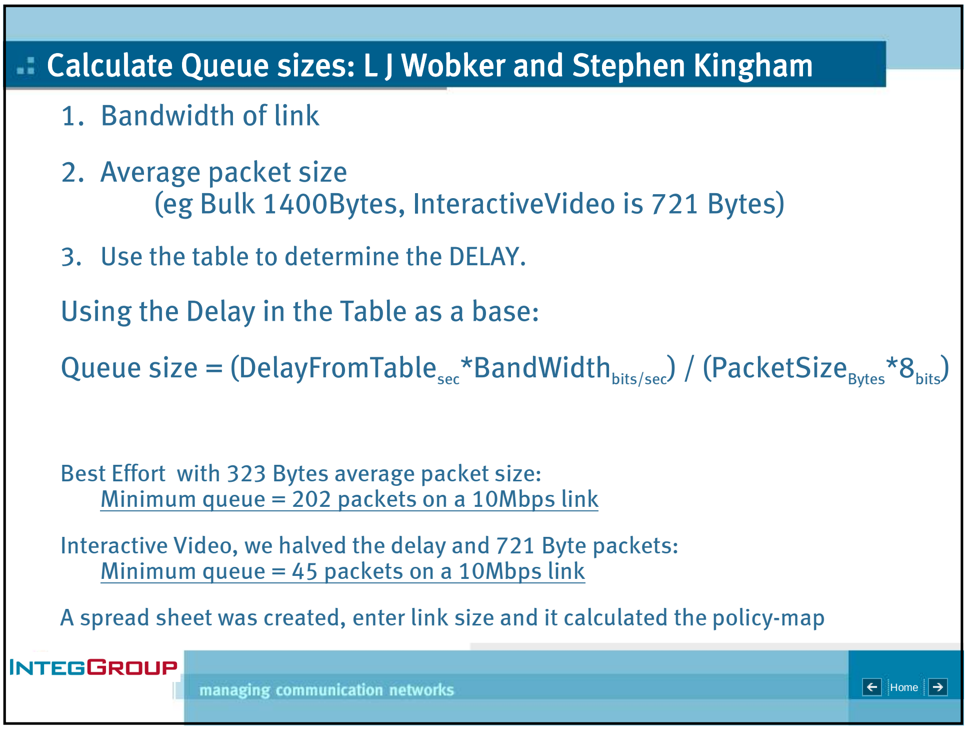
The correct theory and in practice it is not that simple!

Calculate Queue sizes: L J Wobker and Stephen Kingham

Link Speed (bps)	Min Delay (seconds)	Max Delay (seconds)
256,000	0.064	0.160
384,000	0.053	0.160
512,000	0.060	0.160
1,024,000	0.048	0.128
10,000,000	0.052	0.123
34,000,000	0.018	0.054
44,700,000	0.014	0.041
100,000,000	0.010	0.031
1,000,000,000	0.010	0.041

If the delay in the queue reaches the minimum then drop 1 in 10 packets before they leave the router. Do not accept more outgoing packets when delay reaches the maximum, ie drop them all.

The table is derived from a paper "WRED Maximum/Minimum Threshold Recommendations"
by Lawrence J Wobker lwobker@cisco.com, Sep2006



Calculate Queue sizes: L J Wobker and Stephen Kingham

1. Bandwidth of link
2. Average packet size
   (eg Bulk 1400Bytes, InteractiveVideo is 721 Bytes)
3. Use the table to determine the DELAY.

Using the Delay in the Table as a base:

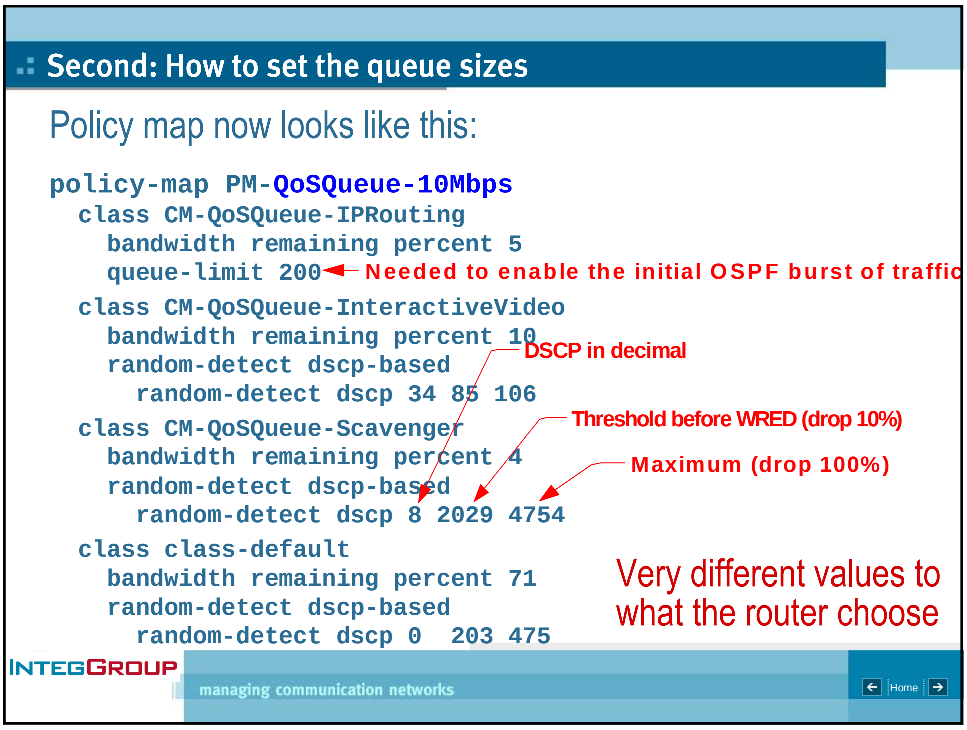
Queue size = $(\text{DelayFromTable}_{sec} * \text{BandWidth}_{bits/sec}) / (\text{PacketSize}_{Bytes} * 8_{bits})$

Best Effort with 323 Bytes average packet size:
Minimum queue = 202 packets on a 10Mbps link

Interactive Video, we halved the delay and 721 Byte packets:
Minimum queue = 45 packets on a 10Mbps link

A spread sheet was created, enter link size and it calculated the policy-map

INTEGGROUP managing communication networks



Second: How to set the queue sizes

Policy map now looks like this:

```
policy-map PM-QoSQueue-10Mbps
  class CM-QoSQueue-IPRouting
    bandwidth remaining percent 5
    queue-limit 200
  class CM-QoSQueue-InteractiveVideo
    bandwidth remaining percent 10
    random-detect dscp-based
      random-detect dscp 34 85 106
  class CM-QoSQueue-Scavenger
    bandwidth remaining percent 4
    random-detect dscp-based
      random-detect dscp 8 2029 4754
  class class-default
    bandwidth remaining percent 71
    random-detect dscp-based
      random-detect dscp 0  203 475
```

Needed to enable the initial OSPF burst of traffic

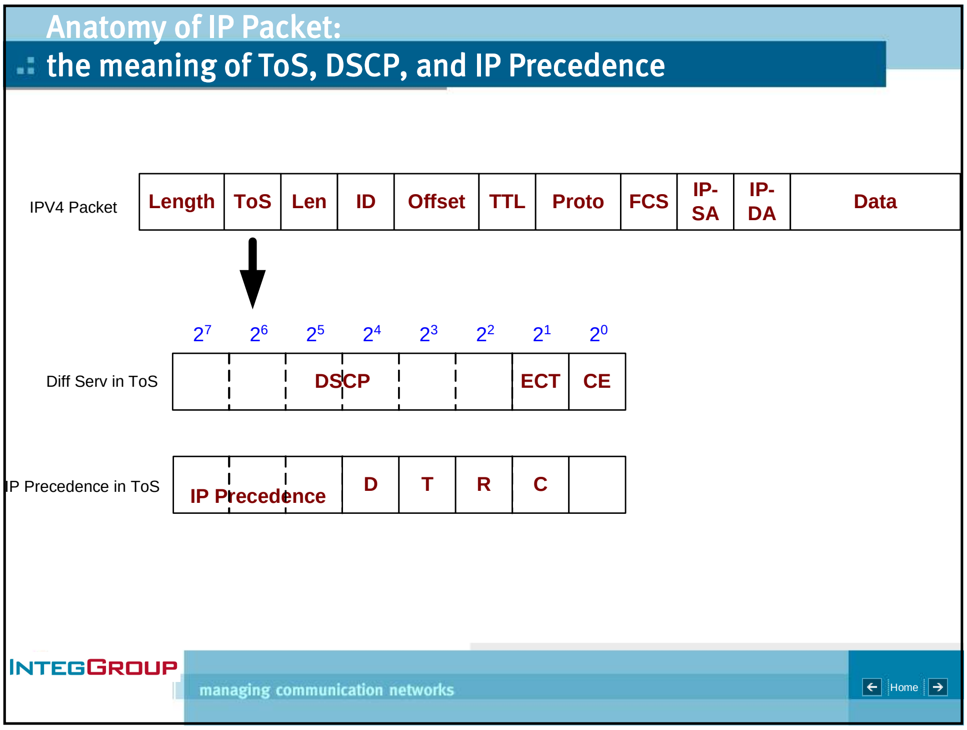
DSCP in decimal

Threshold before WRED (drop 10%)

Maximum (drop 100%)

Very different values to what the router choose

INTEGGROUP managing communication networks



Anatomy of IP Packet: the meaning of ToS, DSCP, and IP Precedence

IPV4 Packet

Length	ToS	Len	ID	Offset	TTL	Proto	FCS	IP-SA	IP-DA	Data

Diff Serv in ToS

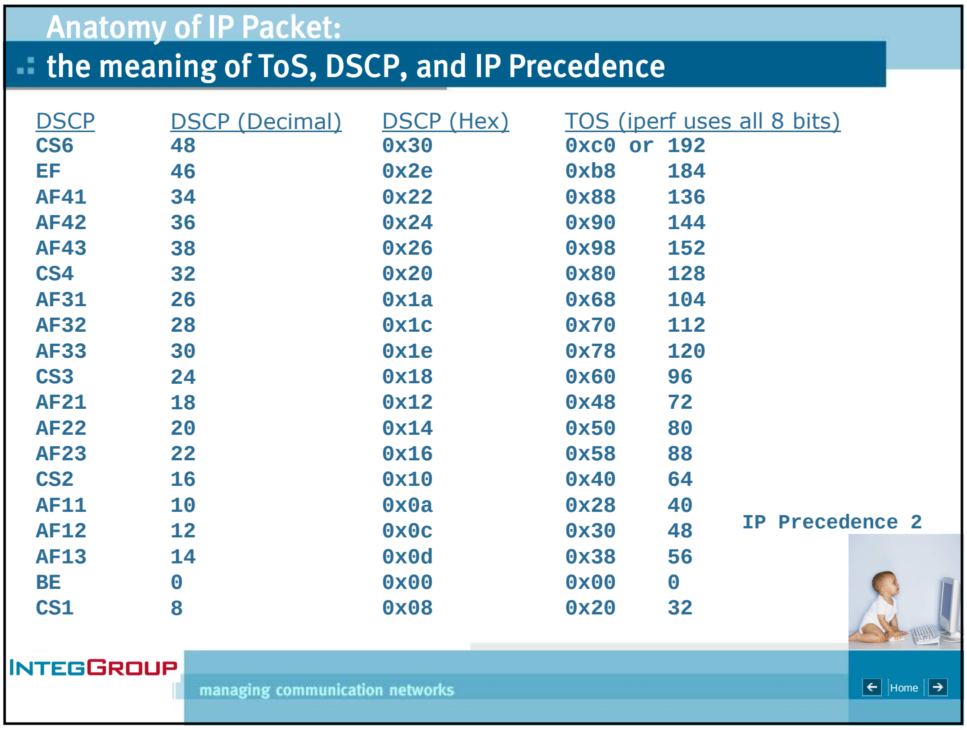
2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
DSCP						ECT	CE

IP Precedence in ToS

2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
IP Precedence			D	T	R	C	

INTEGGROUP managing communication networks

Home



Anatomy of IP Packet: the meaning of ToS, DSCP, and IP Precedence

DSCP	DSCP (Decimal)	DSCP (Hex)	TOS (iperf uses all 8 bits)
CS6	48	0x30	0xc0 or 192
EF	46	0x2e	0xb8 184
AF41	34	0x22	0x88 136
AF42	36	0x24	0x90 144
AF43	38	0x26	0x98 152
CS4	32	0x20	0x80 128
AF31	26	0x1a	0x68 104
AF32	28	0x1c	0x70 112
AF33	30	0x1e	0x78 120
CS3	24	0x18	0x60 96
AF21	18	0x12	0x48 72
AF22	20	0x14	0x50 80
AF23	22	0x16	0x58 88
CS2	16	0x10	0x40 64
AF11	10	0x0a	0x28 40
AF12	12	0x0c	0x30 48
AF13	14	0x0d	0x38 56
BE	0	0x00	0x00 0
CS1	8	0x08	0x20 32

IP Precedence 2

INTEGGROUP managing communication networks

Home

Other active measurement tools in addition to iperf: NDT

Network Diagnostic Tool (NDT)

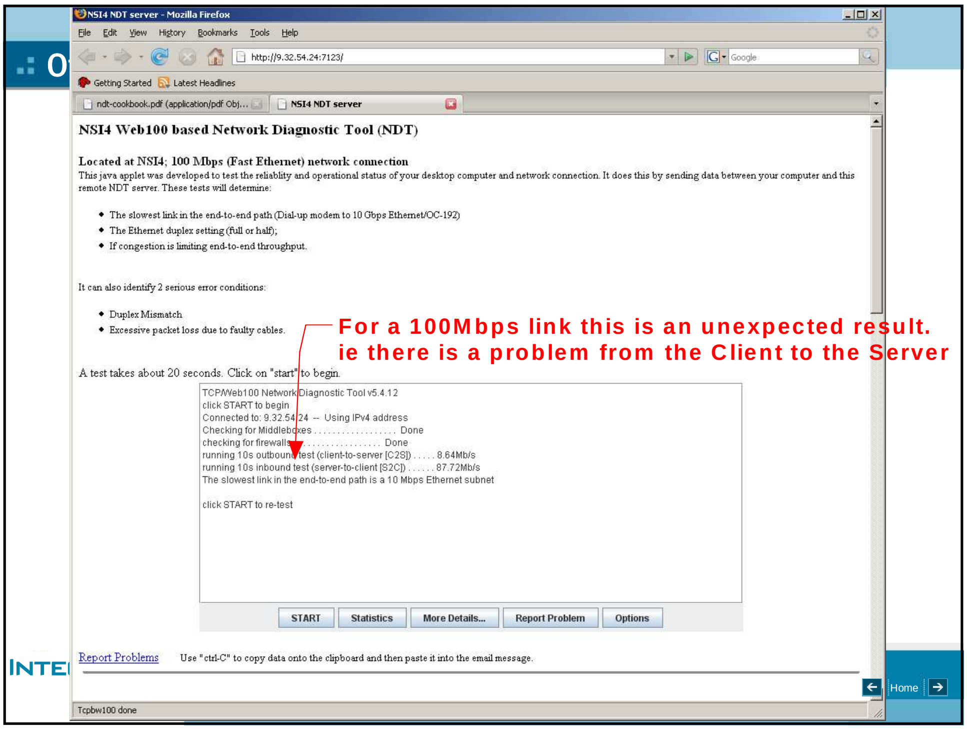
can be run from a client's workstation using their web browser

References:

1. Internet 2 e2epi.internet2.edu/ndt/

2. Build on Fedora Core 9 will be published on www.kingtech.com.au/docs/web100

Used in Centrelink as part of investigating the problem.



NSI4 Web100 based Network Diagnostic Tool (NDT)

Located at NSI4; 100 Mbps (Fast Ethernet) network connection

This java applet was developed to test the reliablity and operational status of your desktop computer and network connection. It does this by sending data between your computer and this remote NDT server. These tests will determine:

- The slowest link in the end-to-end path (Dial-up modem to 10 Gbps Ethernet/OC-192)
- The Ethernet duplex setting (full or half);
- If congestion is limiting end-to-end throughput.

It can also identify 2 serious error conditions:

- Duplex Mismatch
- Excessive packet loss due to faulty cables.

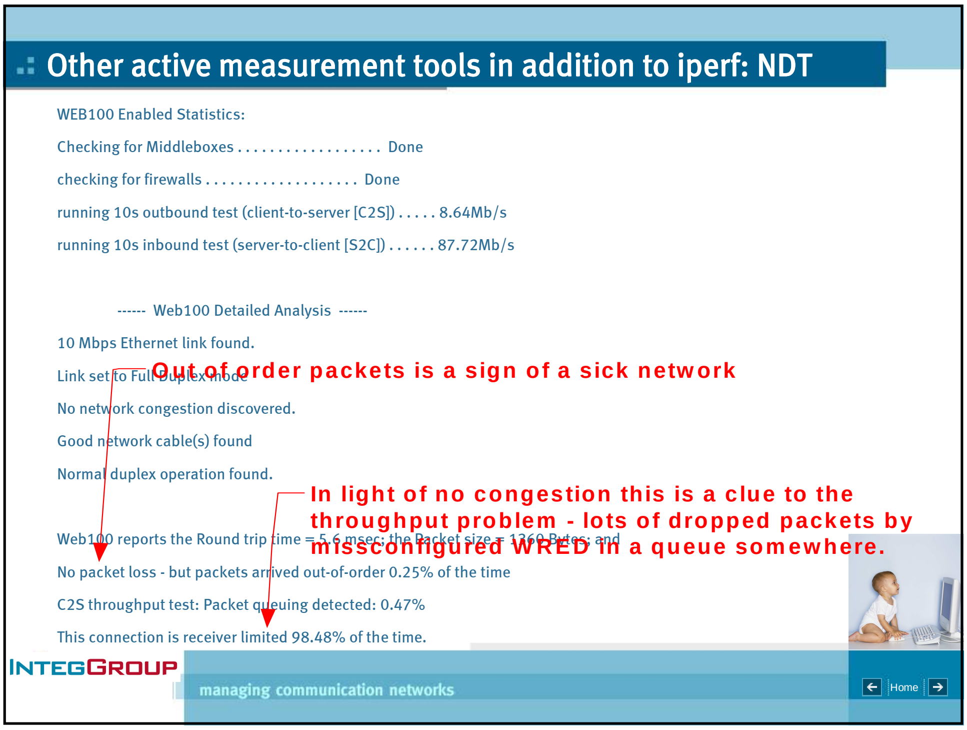
For a 100Mbps link this is an unexpected result.
ie there is a problem from the Client to the Server

A test takes about 20 seconds. Click on "start" to begin.

```
TCP/Web100 Network Diagnostic Tool v5.4.12
click START to begin
Connected to: 9.32.54.24  --  Using IPv4 address
Checking for Middleboxes . . . . . . . . . . . . . . . . . .  Done
checking for firewalls . . . . . . . . . . . . . . . . . . .  Done
running 10s outbound test (client-to-server [C2S]) . . . . . 8.64Mb/s
running 10s inbound test (server-to-client [S2C]) . . . . . . 87.72Mb/s
The slowest link in the end-to-end path is a 10 Mbps Ethernet subnet

click START to re-test
```

Report Problems   Use "ctrl-C" to copy data onto the clipboard and then paste it into the email message.



Other active measurement tools in addition to iperf: NDT

```
WEB100 Enabled Statistics:
Checking for Middleboxes . . . . . . . . . . . . . . . . . .  Done
checking for firewalls . . . . . . . . . . . . . . . . . . .  Done
running 10s outbound test (client-to-server [C2S]) . . . . . 8.64Mb/s
running 10s inbound test (server-to-client [S2C]) . . . . . . 87.72Mb/s

        ------  Web100 Detailed Analysis  ------
10 Mbps Ethernet link found.
Link set to Full Duplex mode
No network congestion discovered.
Good network cable(s) found
Normal duplex operation found.

Web100 reports the Round trip time = 5.6 msec; the Packet size = 1360 Bytes; and
No packet loss - but packets arrived out-of-order 0.25% of the time
C2S throughput test: Packet queuing detected: 0.47%
This connection is receiver limited 98.48% of the time.
```

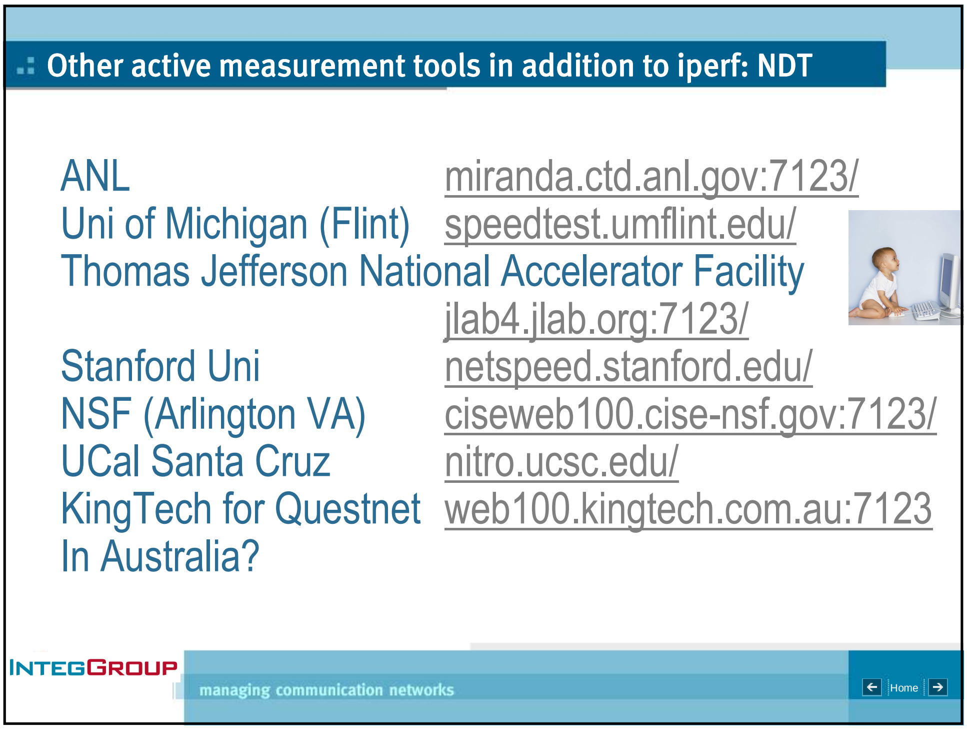
Out of order packets is a sign of a sick network

In light of no congestion this is a clue to the throughput problem - lots of dropped packets by missconfigured WRED in a queue somewhere.

INTEGGROUP

managing communication networks

Home



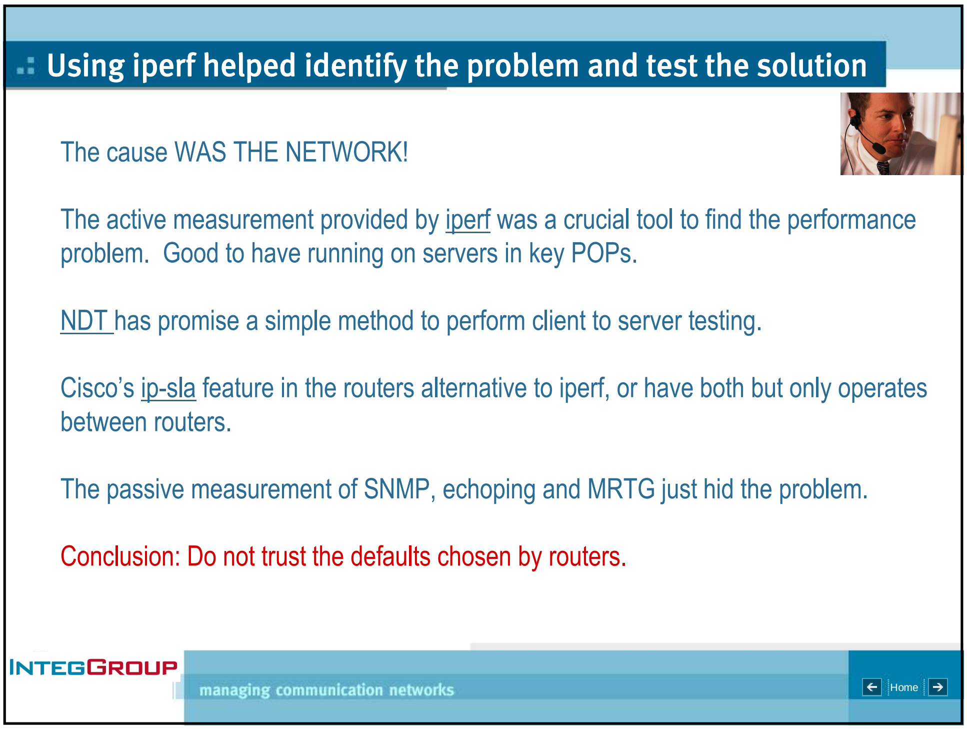
Other active measurement tools in addition to iperf: NDT

ANL	miranda.ctd.anl.gov:7123/
Uni of Michigan (Flint)	speedtest.umflint.edu/
Thomas Jefferson National Accelerator Facility	jlab4.jlab.org:7123/
Stanford Uni	netspeed.stanford.edu/
NSF (Arlington VA)	ciseweb100.cise-nsf.gov:7123/
UCal Santa Cruz	nitro.ucsc.edu/
KingTech for Questnet In Australia?	web100.kingtech.com.au:7123

INTEGGROUP

managing communication networks

Home



Using iperf helped identify the problem and test the solution

The cause WAS THE NETWORK!

The active measurement provided by iperf was a crucial tool to find the performance problem. Good to have running on servers in key POPs.

NDT has promise a simple method to perform client to server testing.

Cisco's ip-sla feature in the routers alternative to iperf, or have both but only operates between routers.

The passive measurement of SNMP, echoping and MRTG just hid the problem.

Conclusion: Do not trust the defaults chosen by routers.

INTEGGROUP

managing communication networks

INTEGGROUP

To reproduce or use this material please refer to author

Thank You

managing communication networks