Path MTU Discovery Blackholes

Phil Dibowitz
University of Southern California

UUASC-LA
December 4, 2003

The Problem

http://bladeforum.wells.org.uk/

“On my sun blade netscape browser, i am not able to access all the websites. Some websites are just fine and others are talking long time and then timeout.”

comp.dcom.sys.cisco

“The client computers at my remote sites can access all but a handful of websites. From the remote routers I can telnet to the website and receive the html document. But, from the client computers (behind those remote routers), I am unable to receive the html document.”

bellsouth.net.support.adsl

“I have the following setup. Machine #1 running XP-Home & SpeedTouch USB DSL modem. Machine #2 running WinME. Machine #3 running XP-Home. All machines network just fine and machine 2 & 3 can get to the interent through machine 1 just fine for about 90% of the websites. However there are a few websites that if accessed through machines 2 or 3 just will not work.”

Before PMTUD

- Client sends a request, including its local MSS.
- Server answers that request with the biggest packet it can fit down the local pipe, or if client MSS is smaller, then it uses that.
- If it can't fit down a pipe along the way, it gets broken into pieces, i.e. fragmented

- Inefficient, slow, CPU intensive

Then came PMTUD

- RFC 1191, November 1990
- Server still responds with the biggest packets both local networks can handle, but sets the DF (*Don't fragment*) bit in the header.
- Intermediary routers will not fragment, but instead send ICMP type 3 code 4, which is *Destination Unreachable: Fragmentation needed, but DF set*.
- The error message can, but does not have to specify maximum size of next hop
- Server resends using size specified, or a smaller size.

Discovering the Discovery

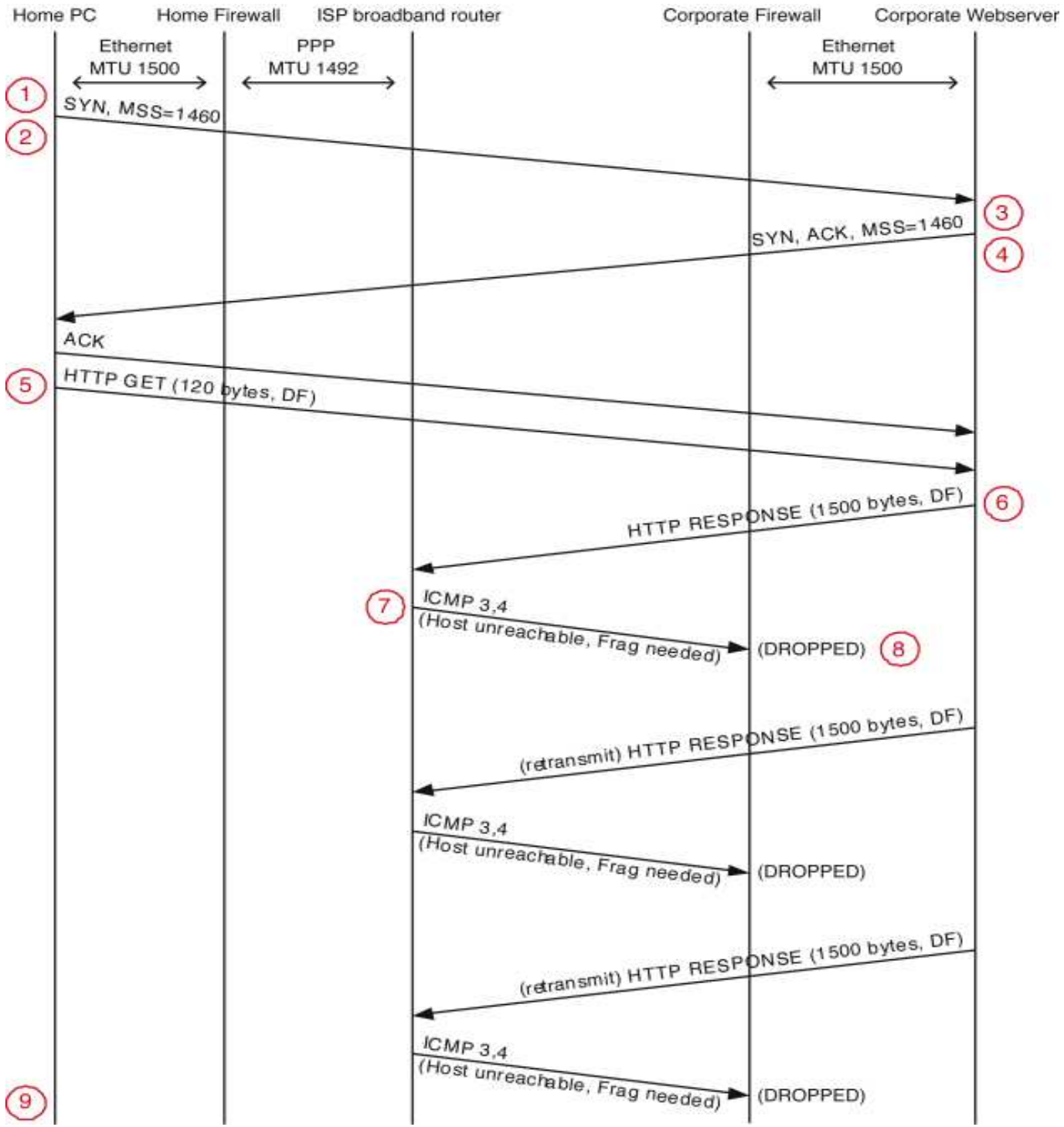
- Allows for near-immediate discovery of maximum end-to-end transmission size.
- ICMP is part of the TCP/IP suite. It always has been. It is reponsible for errors, flow control, and more.
- ICMP is **not** just for pinging.
- While including the size of the next hop is "suggested" and not required, nearly every implimentation known today includes this information to aid in the efficiency of end-to-end communication.

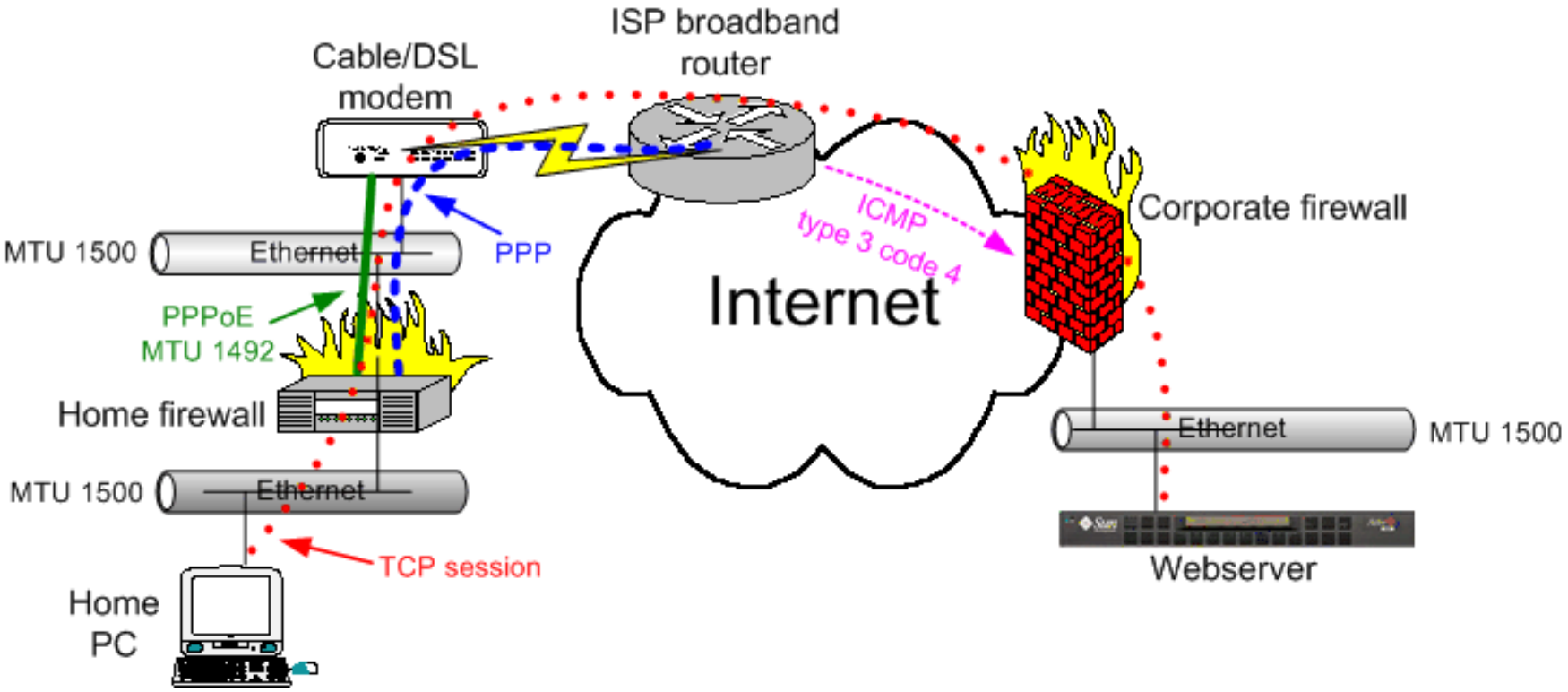
The Blackhole

- A Path MTU Discovery Blackhole occurs when ICMP 3,4 packets do not reach the system that is sending packets that are too large for the smallest MTU on the end-to-end link.
- Causes?
  - Faulty Routers
  - Incorrect Filters
  - Incorrect Firewalls

More on the Blackhole

- So, what happens?
- Server doesn't know to send smaller packets because a firewall in front of it is blocking the packets
- Server retries a few times
- It eventually gives up
- Client times out


Home PC
Home Firewall
ISP broadband router
Corporate Firewall
Corporate Webserver
Ethernet
MTU 1500
PPP
MTU 1492
Ethernet
MTU 1500
1
SYN, MSS=1460
2
3
SYN, ACK, MSS=1460
4
ACK
5
HTTP GET (120 bytes, DF)
6
HTTP RESPONSE (1500 bytes, DF)
7
ICMP 3,4
(Host unreachable, Frag needed)
(DROPPED)
8
(retransmit) HTTP RESPONSE (1500 bytes, DF)
ICMP 3,4
(Host unreachable, Frag needed)
(DROPPED)
(retransmit) HTTP RESPONSE (1500 bytes, DF)
ICMP 3,4
(Host unreachable, Frag needed)
(DROPPED)
9


ISP broadband
router
Cable/DSL
modem
MTU 1500
Ethernet
PPP
PPPoE
MTU 1492
Home firewall
MTU 1500
Ethernet
TCP session
Home
PC
ICMP
type 3 code 4
Internet
Corporate firewall
Ethernet
MTU 1500
Webserver

Some History

- 1988: Path MTU Discovery proposed
- 1991: Finalizes PMTUD, recomends its use to eliminate fragmentation
- 1998: Oldest website we could find mention the blackhole
- 2000: RFC 2923 discusses problems with filtering ICMP and Path MTU Discovery
- 2001: SANS.org: Truth about ICMP
- 2002: The MSS Initiative

More History

- Affects technologies such as X.25, SLIP...
- Small MTUs only at endpoints now, right?

Not Taken Seriously

Prior to the recent growth of broadband....

The number affected was so small, many ignored the problem.

Client-side fixes were considered acceptable.

Recent History

- PPP over Ethernet (PPPoE)
- Point-to-point Tunneling Protocol (PPTP)
- Generic Route Encapsulation (GRE)
- IP version 6 (IPv6)
- 10Gb ethernet
- DSL/cable users on the rise
- Home firewalls

And the problem grows...

- With the use of broadband, and thus these protocols growing fast, many more users are affected.

- More and more questions regarding the blackhole are seen on newgroups and mailing list as time goes on.

Who is (not) affected

1) Just one workstation connected to a modem
2) Home gateways with a public IP address on an Ethernet interface
3) Home gateways connecting to a modem using USB
4) Home gateways connecting to a modem using PPTP
5) Home gateways connecting to a modem using PPPoE

Size of the problem

Sites that really should know better are/were broken:

www.securityfocus.com

www.cert.org

www.verisign.com

www.counterpane.com

www.ntsecurity.com

Solutions

- Allow ICMP 3,4 to reach your servers
- Disable Path MTU Discovery
- Path MTU Discovery Blackhole detection
- Using a Proxy server
- Lowering MTU of local client network
- MSS Clamping

Solution 1: Let PMTUD Work

- Allowing ICMP 3,4 to reach your servers will allow Path MTU Discover to work as it was intended
- SANS has said several times there are no security issues with ICMP 3,4
- If you leave PTMUD on, this is **the only** solution. Otherwise your site is **broken**.

Solution 2: Disable PMTUD

- If you are over paranoid and refuse to allow in ICMP 3,4, then disable PMTUD
- Usually a /proc or sysctl setting
- Will cause fragmentation for paths with small MTUs, but everyone will be able to access your site.
- This is not broken. It perhaps isn't the most efficient in the world, but it's not broken.

Solution 3: Detection

- Many OSes now have detection algorithms
- Must be turned on (/proc, sysctl)
- Might as well just turn PMTUD off
- Slower than PMTUD
- Potentially slightly faster for long-term connections to clients on smaller MTUs

Solution 4: Proxy server

- The client could setup a proxy server.
- Proxy sits outside of small-MTU network, so it is unaffected
- Talks to clients at small MTU,
- Does NOT fix the problem if the small hop is in the middle.
- Client-side fix for server-side problem. Bad.

Solution 5: Lower local MTU

- Could lower MTU of local client network
- Requires change on all machines
- OK for small networks, unacceptable for large ones
- Client side-fix for server-side problem. Bad.
- Will not fix case of a certain path to a certain server has a smaller MTU than you set local MTU to.

MSS Clamping

- An ugly hack
- Border router of the client network changes MSS in packet
- End server sends smaller packets to begin with based on munged MSS.
- Will not fix problems with smaller MTU in the middle
- Client-side fix for server-side problem. Bad.

The MSS Initiative

- Started January 2002
- Contacts administrators of broken sites
- Blacklists sites that don't respond within two weeks (fix not required)
- Offers assistance in correction the problem
- Provides detection instructions for users
- Provides a list of broken sites for comparison for users

The Message

RFC 2923 mentions in Chapter 3:

It is vitally important that those who design and deploy security systems understand the impact of strict filtering on upper-layer protocols. The safest web site in the world is worthless if most TCP implementations cannot transfer data from it.

Conclusion

- Know what you are filtering and why
- Don't assume everything is okay if a simple test scenario seems to work
- Set up and publish technical points of contact
- Listen to your users

Extra Reading

- http://www.ietf.org/rfc/rfc1191.txt
- http://www.ietf.org/rfc/rfc2923.txt
- http://rr.sans.org/threads/ICMP.php
- http://www.cisco.com/warp/public/105/38.shtml
- http://www.phildev.net/mss/

Thanks

- Richard van den Berg
- Rabbs
- All those who have helped out with MSS Initiative
- Anyone who fixes their site! =)