⚙️ TCP Fragmentation

🔷 2. Path MTU and Fragmentation Trigger

- **MTU (Maximum Transmission Unit)**: Largest IP packet size (typically 1500 bytes for Ethernet).
- **MSS (Maximum Segment Size)**: Largest segment TCP can send (MTU - 40 bytes).
- **Fragmentation occurs** if an IP packet exceeds the MTU and:
- PMTUD is disabled or fails
- DF (Don't Fragment) bit is **not set**
- Underlying path contains small-MTU links (VPNs, tunnels, MPLS)

Diagram:

```
App Layer       ————————————————————> Sends 16KB data
TCP Layer       ————————————————————> Breaks into 1460B segments (MSS
IP Layer        ————————————————————> Each segment + headers (~1500B)
↓
Router detects MTU = 1200
→ Fragments each 1500B packet into:
    - Fragment 1: 1200 bytes
    - Fragment 2: 300 bytes
```

🔷 4. Path MTU Discovery (PMTUD)

- TCP uses PMTUD to avoid IP fragmentation.
- Works by sending packets with the **DF** (Don’t Fragment) bit set.
- If a router can’t forward due to MTU, it sends an ICMP type 3 code 4 message: **"Fragmentation needed and DF set"**

PMTUD Failure: "Black Hole" Symptoms

- ICMP blocked by firewall (common in enterprises)
- Large packets mysteriously disappear
- Connections hang during TLS handshakes, file transfers, etc.

🔷 6. Fragmentation and TCP Performance Tuning

🔶 Tunneling Protocols

- IP-in-IP, GRE, IPSec add 20-60 bytes of headers.
- Reduces effective MTU, often to **\~1400 or lower**.
- Neglecting this causes fragmentation or black holes.

🔶 MSS Clamping

Ensures TCP doesn't try to send segments larger than the path can carry:

```
# Example: Clamp MSS to 1360
iptables -t mangle -A FORWARD -p tcp --tcp-flags SYN,RST SYN \
         -j TCPMSS --clamp-mss-to-pmtu
```

> MSS clamping is especially important on **edge routers**, **VPN concentrators**, and **cloud VPC gateways**.

🔷 8. Kernel Internals and Fragmentation

Fragmentation Logic in Linux

- Linux fragments **before** queuing to NIC.
- Uses **skb (socket buffer)** structure.
- Fragment queues tracked via `/proc/net/ip_frag` (on older kernels).
- Reassembly timeout is \~30 seconds (`/proc/sys/net/ipv4/ipfrag_time`)

Netfilter Hooks

- Fragmented packets hit the `PREROUTING` chain.
- Fragments do **not pass through** `INPUT` in the same form.
- IDS must reassemble fragments **before** TCP stream reassembly.

✅ Best Practices Summary

Area	Recommendation
Security	Drop or limit fragmented traffic unless necessary
Performance	Enable MSS clamping on tunnels, VPNs
Debugging	Use `ping -M do`, `tcpdump`, `traceroute -F`
PMTUD	Allow ICMP type 3 code 4 messages through firewalls
IPv6	Don’t rely on fragmentation; use PMTUD or PLPMTUD

📌 Real-World Scenarios

1. **VPN tunnel breaks large TCP transfers** → Fix: Clamp MSS to \~1350
2. **TLS handshake fails intermittently** → Root cause: PMTUD broken, fragmentation blocked by intermediate device
3. **IDS/IPS fails to detect attack** → Attacker used fragmented payloads to evade deep packet inspection