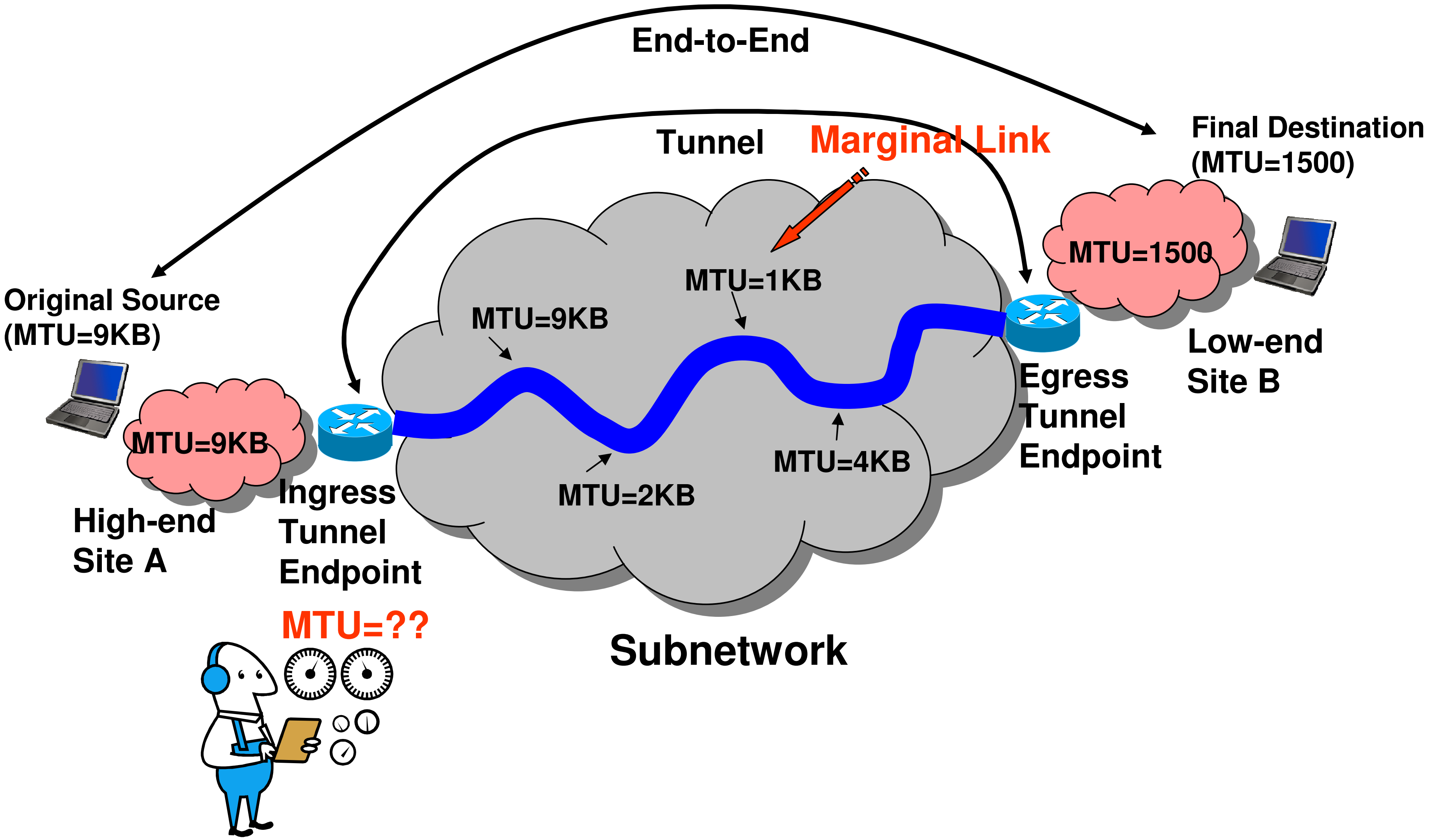
Subnetwork Encapsulation and Adaptation Layer (SEAL)

IETF76 INTAREA Meeting

Fred L. Templin
fred.l.templin@boeing.com

Tunnel Maximum Transmission Unit (MTU)



SEAL Approach

- **Used with IP-in-IP encapsulation**
- **4Byte encapsulation sublayer**
- **Each packet has a 32bit sequence number**
- Track MTU **w/o classical path MTU discovery**
- **Detect** and **tune out** in-the-network IPv4 fragmentation
- Segmentation to mitigate **misconfigured MTUs** and **marginal links**
- Promotes desired end-state of **MTU-robust Internet**

- **Works just like IPv6 fragmentation, except:**
  - **fixed segment size**
  - **non-overlapping segments**
  - **ETE informs ITE of Maximum Receive Unit (MRU)**
  - **no prior negotiations between ITE and ETE needed**

Draft Status

- Significant improvements based on list review input
- Standards-track submission through INTAREA
- Two distinct “modes” of operation:
    - SEAL-FS (SEAL with Fragmentation Sensing)
        - used when all links in the network have MTU of at least M (e.g., 1500)
        - ETE senses IPv4 fragmentation; sends report to ITE
    - SEAL-SR (SEAL with Segmentation and Reassembly)
        - used when end systems need to see an assured MTU of at least M
        - used when end systems prefer a larger MTU
        - ETE senses IPv4 fragmentation; sends report to ITE
        - ITE segments large packets; ETE reassembles

SEAL With Fragmentation Sensing (SEAL-FS)

- Minimal mechanism for discovering tunnel MTU
- Egress Tunnel Endpoint (ETE):
  - Informs ITE of MRU without need for pre-negotiations
  - listens for IP fragmentation and drops all IP fragments
  - sends “Fragmentation Reports” to Ingress Tunnel Endpoint (ITE)
- ITE adjusts tunnel MTU based on fragmentation reports
- ITE never has to segment and ETE never has to reassemble
- Use cases:
  - performance-intensive core routers that support many tunnels over paths containing robust links (MTU >> 1500)

SEAL With Segmentation and Reassembly (SEAL-SR)

- Same as SEAL-FS, but also includes segmentation and reassembly at a layer below IP
- **MTU based on maximum size the ETE can reassemble**; NOT on the link with the smallest MTU in the path
- **End systems see a solid minimum MTU (e.g., 1500),** and can often send packets that are larger than the actual path MTU
- **Supports IPv6 jumbograms even if not all links in the path support jumbograms**
- **Treats reassembly timeouts as indication to reduce MTU**
- Use cases:
  - Enterprise routers connecting high-performance data centers
  - CPE routers
  - MANET routers

Observations

- **“Unmitigated** Fragmentation Considered Harmful”
- **“Carefully-managed** Fragmentation Considered **Useful”**
- In-the-network fragmentation as **“canary in the coal mine”**

For more information:

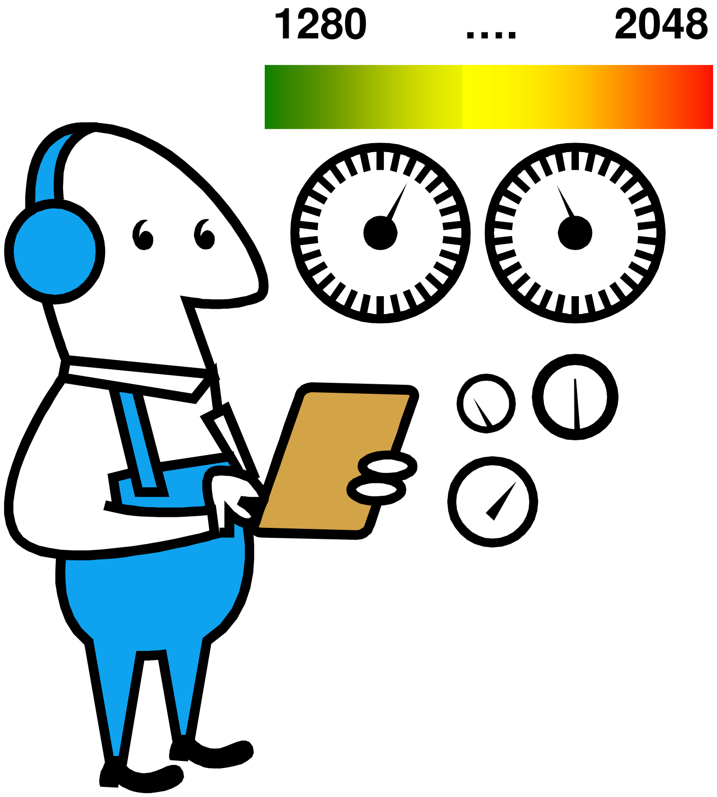
http://tools.ietf.org/html/draft-templin-intarea-seal (specification)
http://osprey67.com/seal (linux source code)

BACKUPS

Problems with Classical Path MTU Discovery

- ICMPs may be lost, erroneous, fabricated
- ICMPs may have insufficient information for relaying
- ALWAYS drops packets when MTU insufficient
- In-the-network tunnels may have 1000's of packets in-flight when a routing change hits an MTU restriction:
  - all packets are dropped
  - flood of ICMPs returned to ITR
  - resources wasted

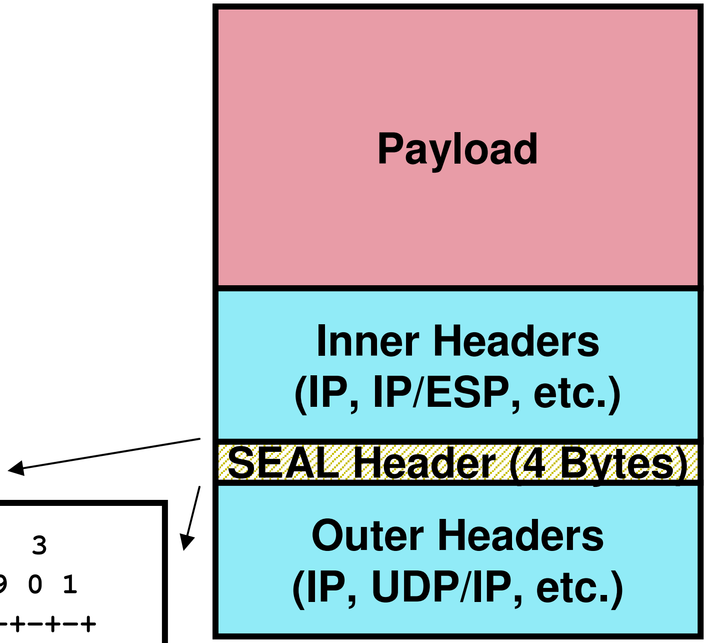
MTU Configuration Knob



- < 1280: MinMTU underflow
- < 1400: fragmentation unlikely
- < 2048: fragmentation managed
- 2048 – 64KB: best-effort
- > 64KB: jumbogram

SEAL Encapsulation

- Extends IP-ID to 32 bits
- Report Fragmentation mechanism
- Tunnel segmentation and reassembly
- Nonce-protected error feedback
- Compatible with wide variety of tunnels



```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |          ID Extension         |A|R|M|RSV| SEG |  Next  Header  |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+


      ID Extension (16 bits)

      A - Acknowledgement Requested (1 bit)

      R - Report Fragmentation (1 bit)

      M - More Segments (1 bit)

      RSV - Reserved (2 bits)

      SEG - Segment number (3 bits)

      Next Header (8 bits)
```