IP-TFS

Christian Hopps

LabN Consulting, LLC

Key Design Points

- Improves on existing IPsec solution
- Unidirectional (like IPsec)
- Constant send rate, Fixed packet size
- Congestion Controlled and Non-CC operating modes
- Uses IPsec
- [Optional] Uses IKEv2
- Minimize configuration required

Unidirectional/Bidirectional

- Data path is unidirectional
  - Sender to Receiver
- Congestion-Control (CC) info is sent in reverse direction
  - Receiver to Sender
- Configure 2 paths for bidirectional operation

Constant Send Rate/Fixed Packet Size

- Packet size never varies
- Packet size manual or automatic configuration
- Use Path MTU Discovery for automatic optimal configuration
- Constant send rate
- Provides requisite transport flow confidentiality

Variation Fully Allowed

- Egress must accept packets at any rate.
- Egress must accept packets of any size.
- IPSec tunnels can start in normal "IP Mode" and transition to IP-TFS.
  - SA reset required to leave IP-TFS mode.

Congestion Controlled (CC) Mode

- Send rate (PPS) adjusted, packet size fixed
  - Congestion causes packet drops not byte drops
- CC Info sent from Receiver to Sender in IKEv2 notification data.
- Sender uses congestion control algorithms to modify packet send rate.
- CC algorithm a local choice
  - No need to standardize
- Circuit breaking supported.
- ECN supported but off by default.

Non-Congestion-Controlled Mode

- For use when IP path bandwidth can be guaranteed.
- Packet loss reported by receiver to admin/operations.
- Optional CC info can be used to report packet loss from sender.
- Optional CC info can be used for circuit breaker.

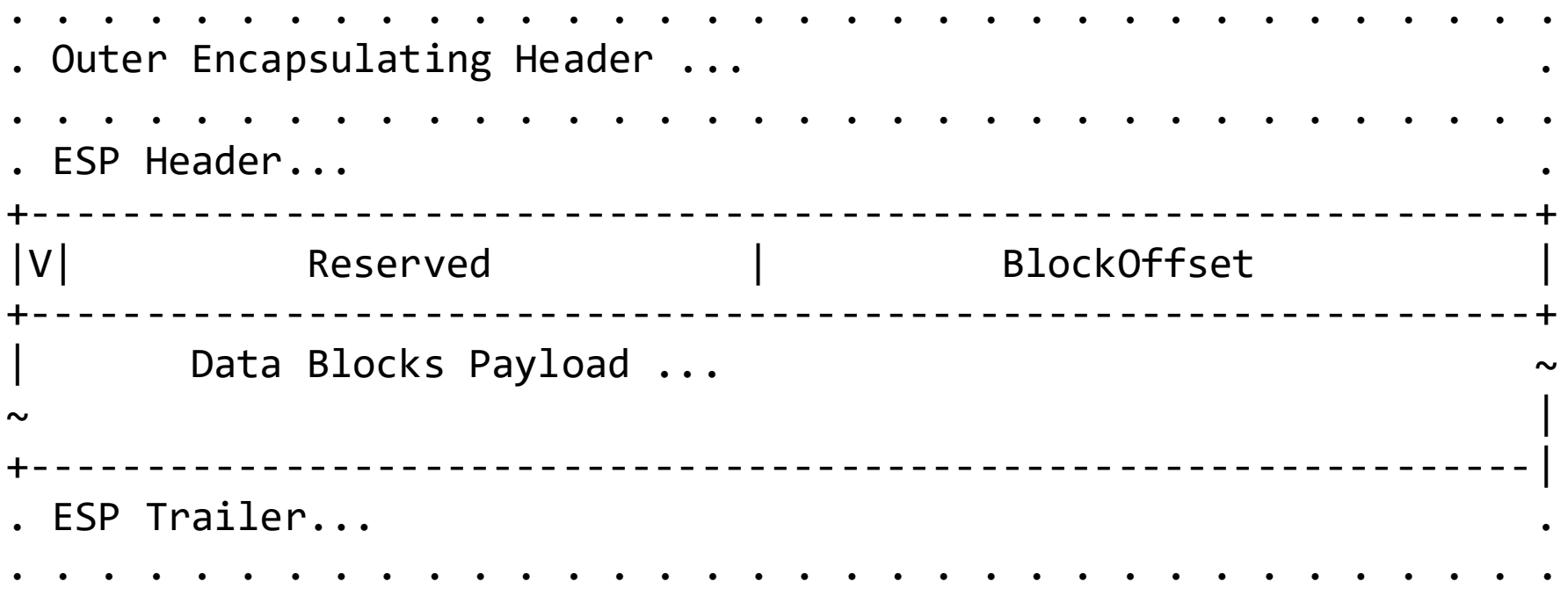
IPSec Transport

- Use IPSec/ESP (encrypted encapsulation) as transport
- Input packets are fragmented and aggregated into IPSec/ESP packet stream
- New IP Protocol Number for new ESP payload (framing)

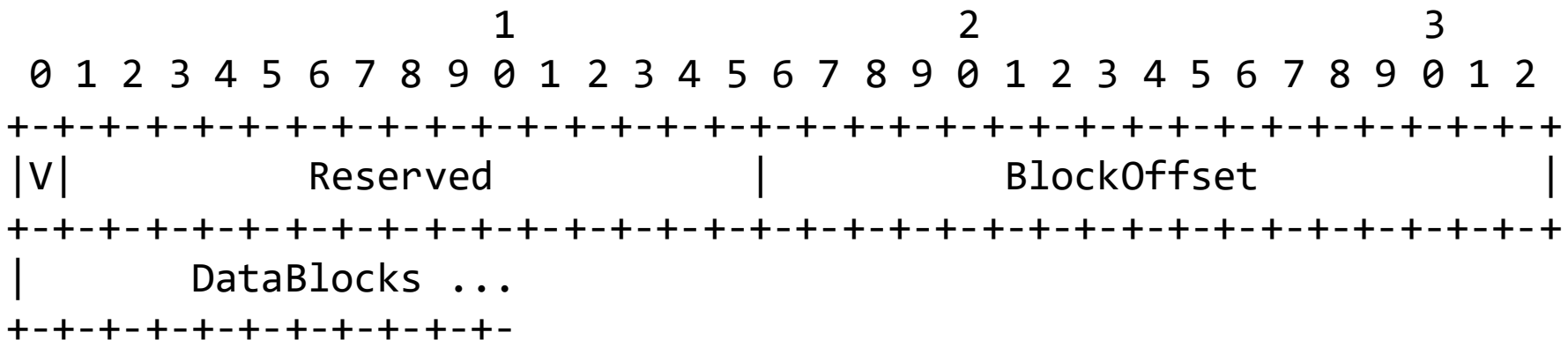
IKEv2 (CC Info)

- Use IKEv2 for CC info advertisement
- Use INFORMATION "exchange" Notification Data
- Periodic send interval (e.g., 1 per second)
- Non-reliable transport (I.e., not a REQUEST/RESPONSE exchange)
- CFG_REQUEST/CFG_REPSPONSE used to configure interval.
- 0 interval allowed for no send.

IP-TFS Packet Format

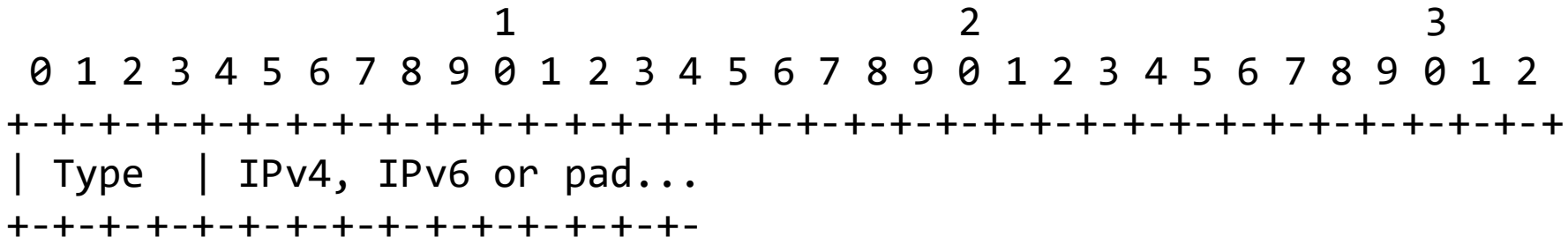


ESP Payload Format



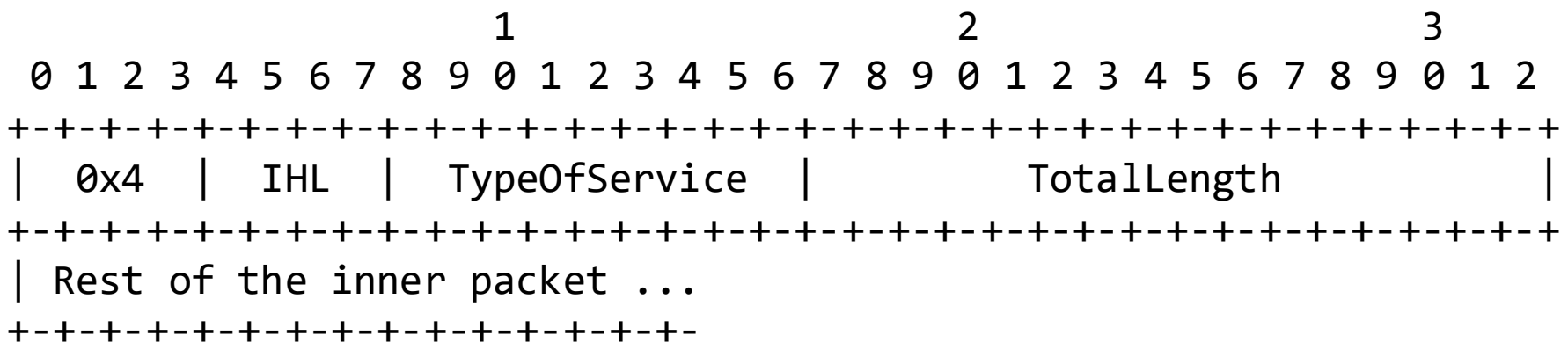
- **V** :: Version, must be set to zero and dropped if set to 1.
- **Reserved** :: set to 0 ignored on receipt.
- **Block Offset** :: This is the number of bytes before the next IP/IPv6 data block. It can point past the end of the containing packet in which case this packet is the continuation of a preivous one and possibly padding. NOTE: This can point into the next packet and yet the current packet can end with padding. This will happen if there's not enough bytes to start a new inner packet in the current outer packet.
- **Data Blocks** :: variable number of bytes that constitute the start or continuation of a previous data block.

Data Blocks



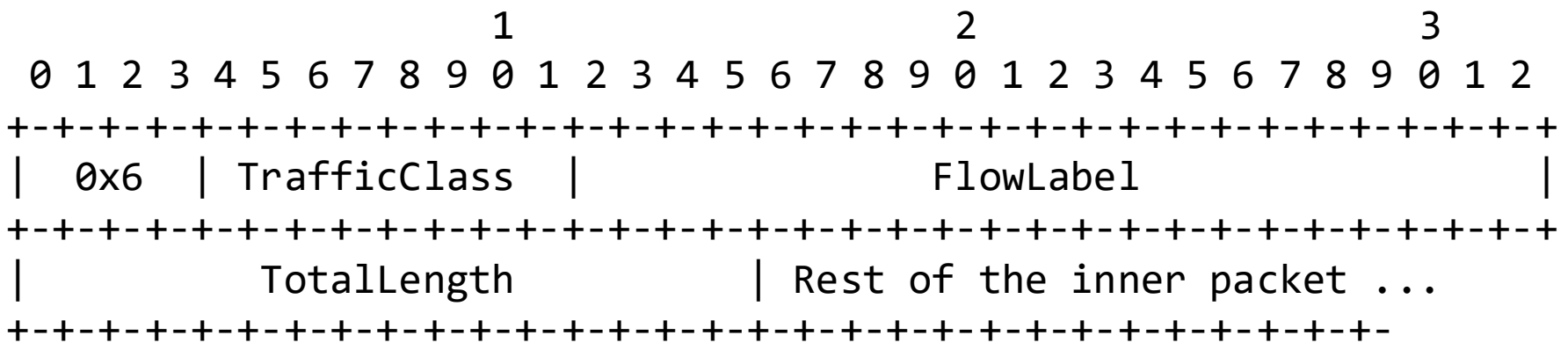
- **Version**
  - 0x0 for pad
  - 0x4 for IPv4
  - 0x6 for IPv6.

IPv4 Data Blocks



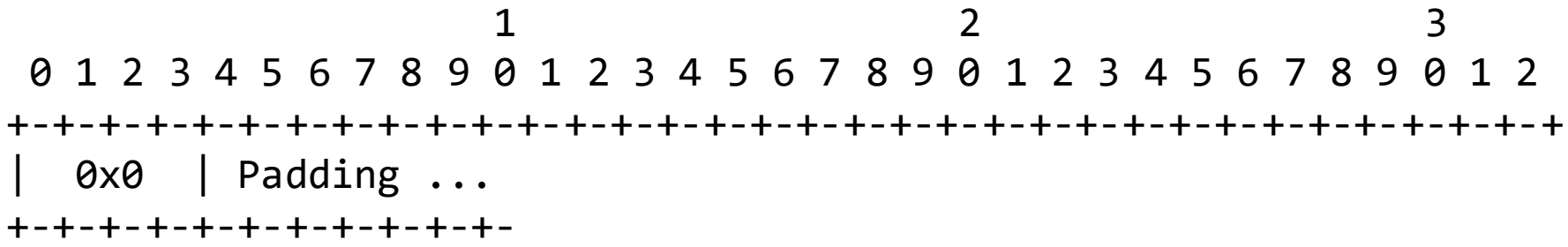
- **Version** :: 0x4 for IPv4
- **Total Length** :: Length of the IPv4 inner packet.

IPv6 Data Blocks



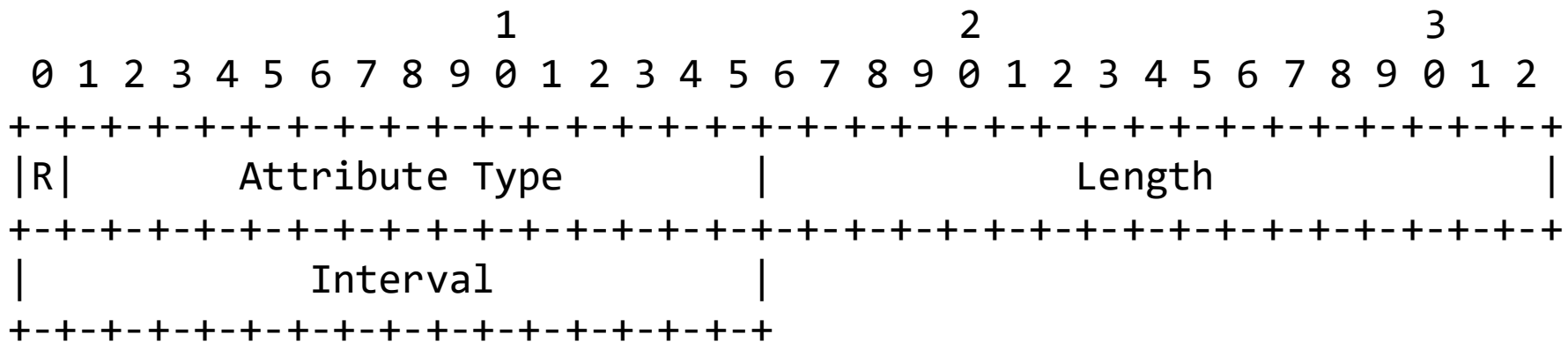
- **Version** :: 0x6 for IPv6
- **Total Length** :: Length of the IPv4 inner packet.

Pad Data Blocks



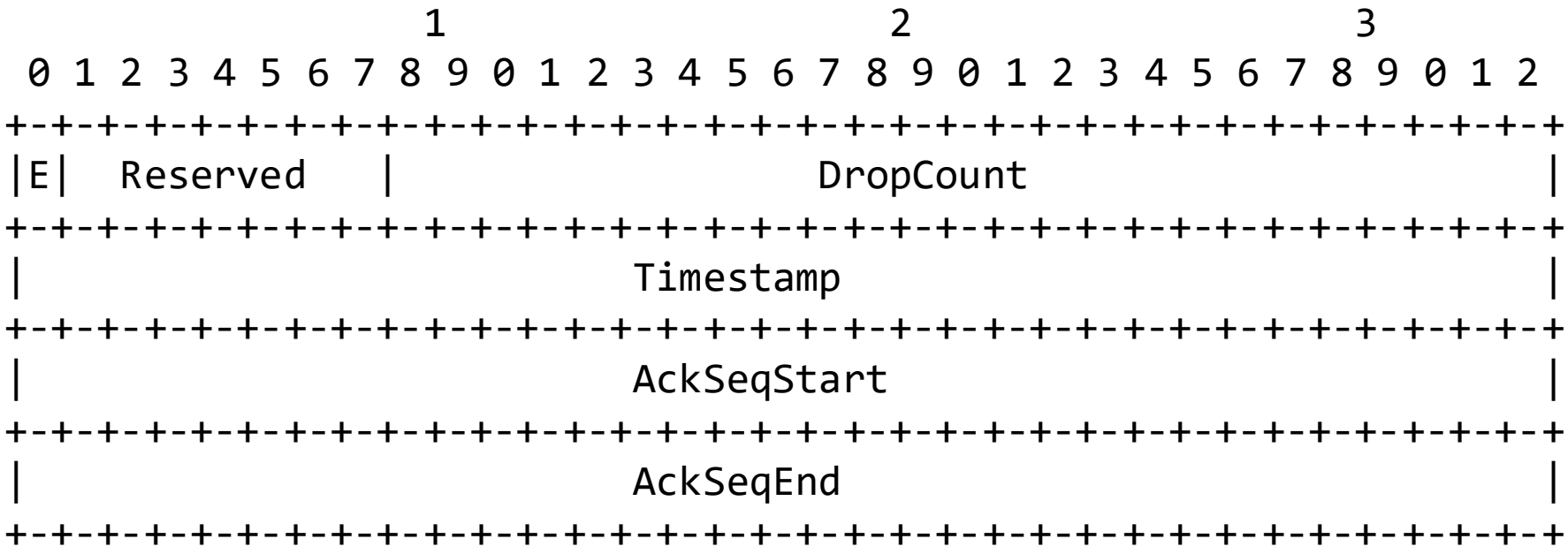
- **Version** :: 0x0 for Padding
- **Padding** :: extends to end of the encapsulating packet.

IKEv2 Config CC Info Sending Interval Attribute



- **R**:: 1 bit set to 0
- **Attribute Type**:: 15 bit value set to TFS_INFO_INTERVAL (TBD).
- **Length**:: 2 octet length set to 2.
- **Interval** :: 2 octet unsigned integer. The sending interval in milliseconds.

CC Info Notification Data



- **E** :: A 1 bit value that if set indicates that packet[s] with Congestion Experienced (CE) ECN bits set were received and used in calculating the DropCount value.
- **Reserved** :: set to 0 ignored on receipt.
- **DropCount**:: For ack data block this is the drop count between AckSeqStart and AckSeqEnd, If the drops exceed the resolution of the counter then set to the max value.
- **Timestamp** :: The time when this notification was created and sent.
- **AckESPSeqStart** :: The first ESP Seq. Num. of the range that this information relates to.
- **AckESPSeqEnd** :: The last ESP Seq. Num. of the range that this information relates to.

Backup Slides

Overhead of bytes per inner packet.

Type	IPSec	IPSec	IPSec	TFS	TFS	TFS
PktSize	576	1500	9000	576	1500	9000
64	512	1436	8936	4.8	1.8	0.3
128	448	1372	8872	9.6	3.5	0.6
256	320	1244	8744	19.1	7.0	1.1
536	40	964	8464	40.0	14.7	2.4
576	600	924	8424	43.0	15.8	2.6
1460	312	40	7540	109.0	40.0	6.5
1500	272	1524	7500	111.9	41.1	6.7
8960	560	144	40	668.7	245.5	40.0
9000	520	1624	9024	671.6	246.6	40.2

Overhead as percentage of inner packet.

Type MTU	IPSec 576	IPSec 1500	IPSec 9000	TFS 576	TFS 1500	TFS 9000
64	800.0%	2243.8%	13962.5%	7.5%	2.7%	0.4%
128	350.0%	1071.9%	6931.2%	7.5%	2.7%	0.4%
256	125.0%	485.9%	3415.6%	7.5%	2.7%	0.4%
536	7.5%	179.9%	1579.1%	7.5%	2.7%	0.4%
576	104.2%	160.4%	1462.5%	7.5%	2.7%	0.4%
1460	21.4%	2.7%	516.4%	7.5%	2.7%	0.4%
1500	18.1%	101.6%	500.0%	7.5%	2.7%	0.4%
8960	6.2%	1.6%	0.4%	7.5%	2.7%	0.4%
9000	5.8%	18.0%	100.3%	7.5%	2.7%	0.4%

Questions and Comments

- More pictures of aggregated inner packets in outer.