Datagram Forwarding via Stateless Internetwork Switching

Gregory G. Finn, Rodney Van Meter, Craig Milo Rogers

NetWorld+Interop '97 Engineer Conference
Las Vegas, May 1997

Switched IP Forwarding

IP forwarding speeds today are too slow for fiber-optic channels.

Generic proposed fix:

- Replace hop-by-hop IP routing with switched forwarding
- Make IP forwarding purely a hardware function

Switching requires forwarding state that is contained in:

- Each packet and switch along a packet's path, or
- Entirely in each packet **(source route)**

Switched IP Forwarding

Common approach locates forwarding state in switches & packets.

- State is distributed to each switch along a route
- Distributed state tables are maintained
- State table limits # distinct routes through a switch

System-area approach puts all forwarding state in each packet.

- Complete forwarding state is located at source
- No distributed state tables to maintain
- Supports indefinite # of distinct routes through a switch
- Switch is independent and passive

OvIP

OvIP (Overpass IP) applies source-routed, system-area switching methodology to the problem of IP forwarding.

Motivations:

- Simple: No distributed forwarding state to maintain
- Faster: No table lookup in switch
- Attractive scaling: No switch table-size limitations
- Not dependent on one link-layer **(Internet, end-to-end)**

Remainder of presentation:

- Brief overview of OvIP
- Problems merging switching and IP methodologies
- OvIP overheads

Switched Internet Forwarding

Switch State	Link-layer specific	Not link-layer specific
distributed flow state	ATM + signalling IP Switching	Tag Switching
no distributed flow state	SANs	

OvIP (Overpass IP) fills in that blank.

End-to-End Switching

Ideal end-to-end switching methodology would be:

- Simple
- Small cost to routers
- Support indefinite # of flows through switches
- Intranet & Internet compatible **(link-layer independent)**

Basic OvIP Approach

Gather hop-by-hop switching state inside a source route (SR). Prefix outgoing datagrams with their SR.

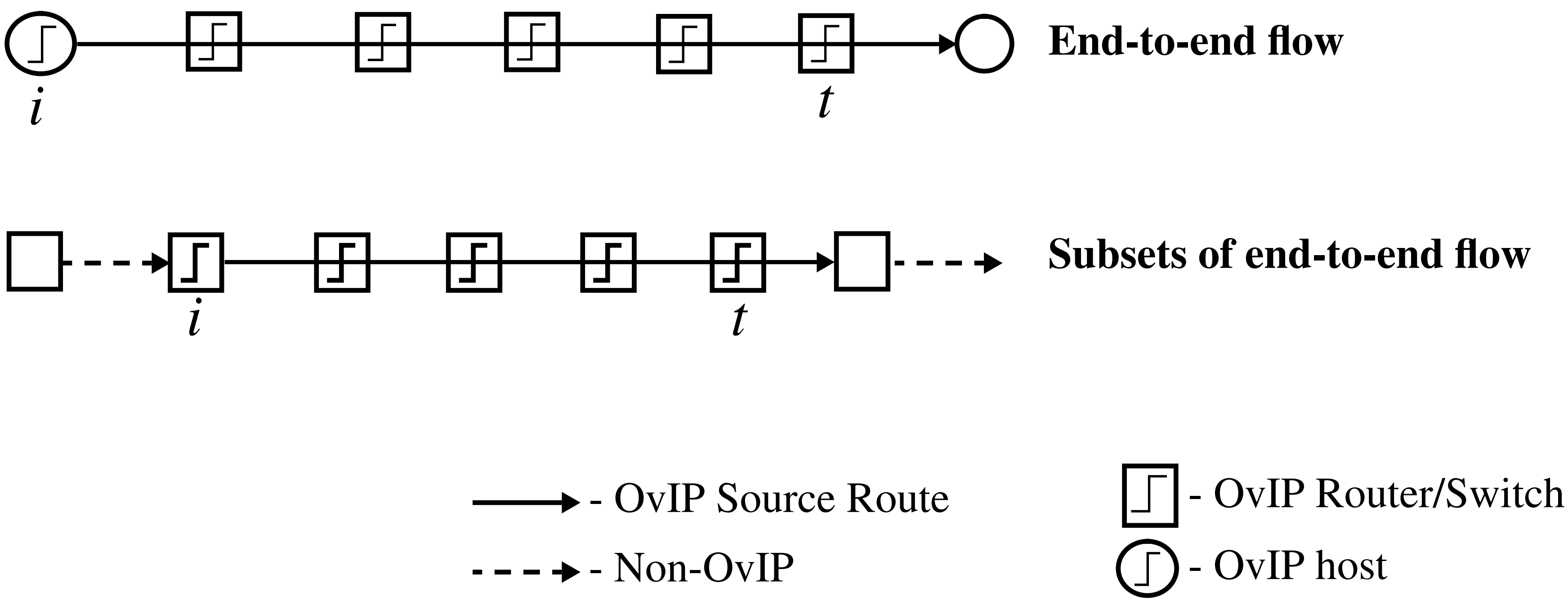
- Per-hop state is privately defined by each router/switch
- Determines outgoing link and link-layer header

OvIP switching methodology is characterized by:

- Majority of mechanism at source **(timers and tables)**
- Small cost to routers **(< 1% overhead packets)**
- Stateless, network-layer switching **(no distributed flow state)**

OvIP

SR extends from *initiator i* to *terminator t*.



Deriving OvIP Source Routes

One method: Analog of *traceroute()* -- Route Ping.

- Message implemented by OvIP router/switches
- Gathers SR hop-by-hop and returns it

But underlying topology changes break OvIP SRs.

- Must not cause connections to break
- Periodically reissue Route Pings (RPs) to refresh SR
- Too much overhead if topology changes frequently

OvIP assumes source routes are relatively stable. Are they?

Source Route Stability

Internet routes are relatively stable.

- 99% end-to-end stable ≥ 10 min. (Paxson)
- < 1% vary on order of minute or less

Fluttering and deflection routing policies produce most of < 1%. Don't build SRs across these.

SR Refresh

To prevent broken connections ---

- Initiators reissue RPs to keep SRs current
- This overhead = #RPs / # packets switched per flow

Cost of maintaining SR depends on refresh period.

- Connection timeouts **(>25 seconds)**
- For initiator $\boldsymbol{i}$, the reissue period $\boldsymbol{PR_i} \leq 12$ seconds

Dynamic OvIP Overhead

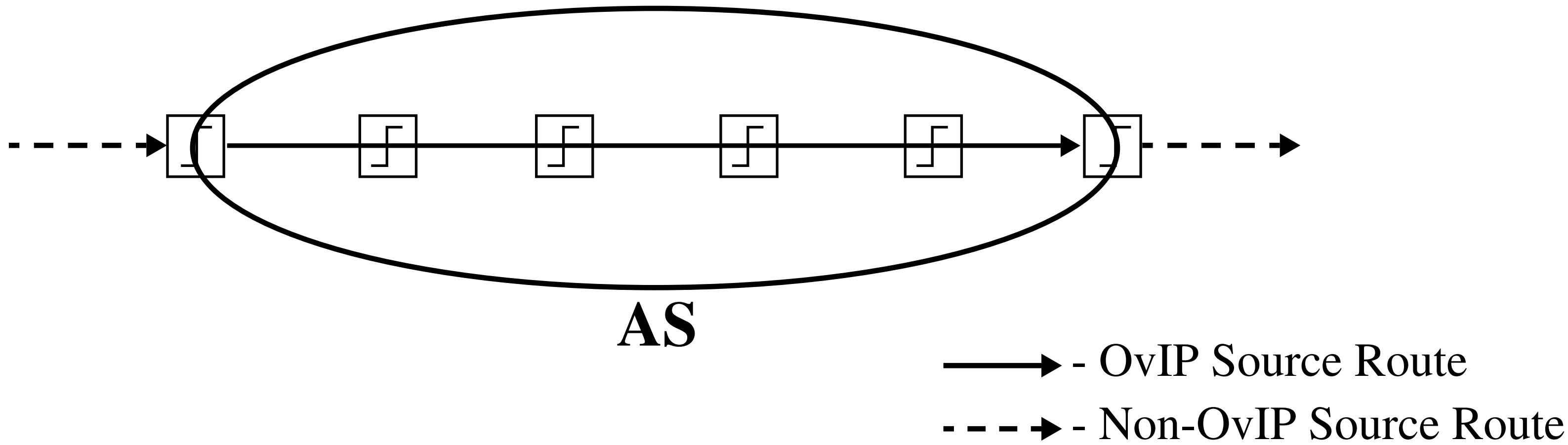
Overhead proportional to ratio of PR_i to flow rate.

- **F(*i*, *dst*)** -- pkt/sec for flow **F** and initiator ***i*.**
- Assume $\boldsymbol{PR_i} \cong 0.1$ sec
- **~** $\boldsymbol{PR_i}$ **/** F(***i***, ***dst***) $\cong$ 0.1 sec **/ F(*i*, *dst*)**

For F(*i*, *dst*) $\geq$ 10/sec, overhead $\leq 1\%$.

For an MPEG-2 flow (≥ 4 Mb/s, 1500 byte/packet), overhead $\leq 0.03\%$.

Overhead Depends Upon Implementation Model



Within AS maintain SRs between border router/switches.

- Aggregate all transit traffic
- **F(*entry*, *exit*)** >> 10 is likely
- $\Rightarrow$ ***PR$_i$*** **/ F(*entry*, *exit*)** very small

Static OvIP Overhead

Two principal sources.

- Increased packet length
- Initiator forwarding tables

SR length ≤ (diameter $*$ per-hop field size)

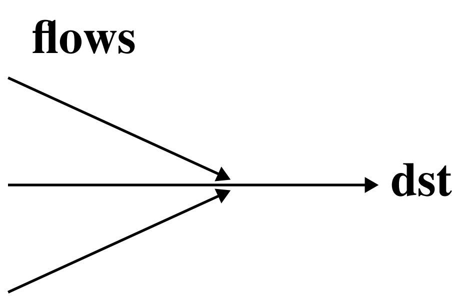
- 30 hops * 2 = 60 octets
- SR consumed during transit ⇒ ~ SR length / 2

Initiator controls static and dynamic overhead.

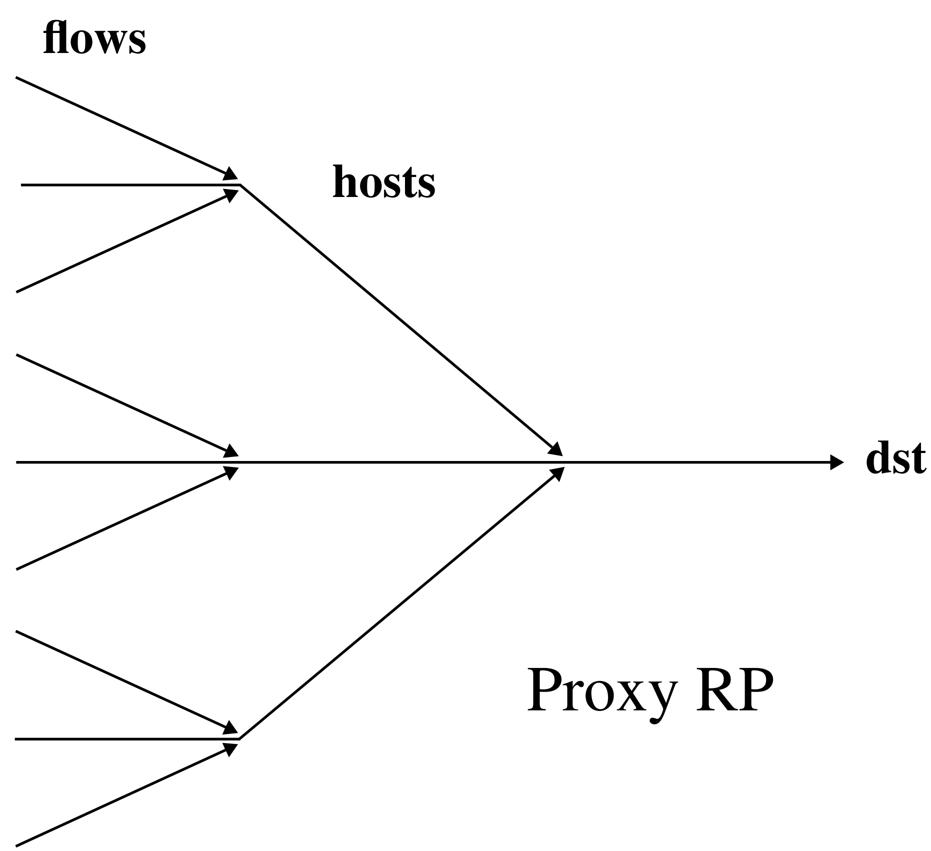
- Use OvIP only when distance ≥ hop count threshold
- Aggregate flows by destination mask

Source Aggregation

Source Host

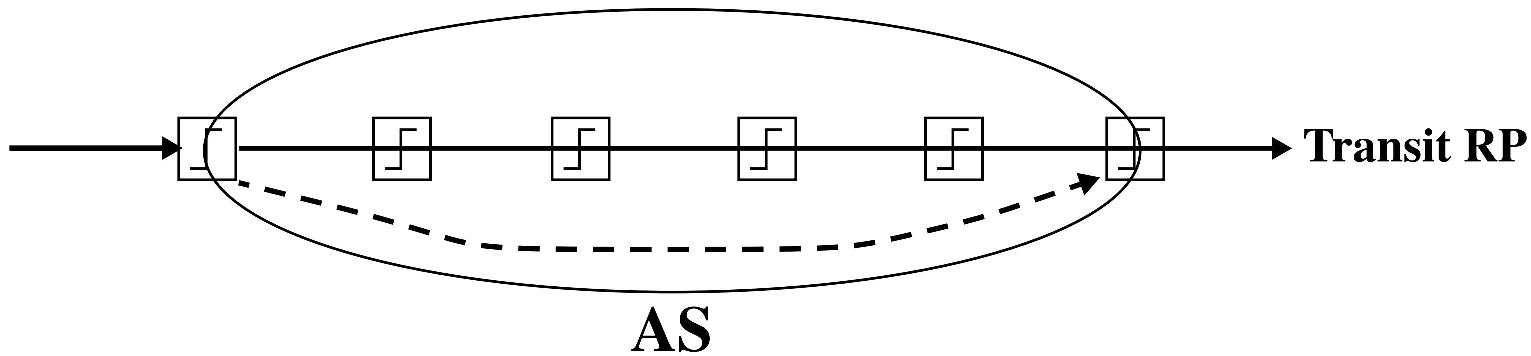


Source Gateway



Further Overhead Reductions

Reduce RP overhead when transiting AS.



Border appends transit SR portion, tunnels RP to exit.

Pros and cons:

- RPs processed reduced from AS Diameter to 2
- Border routers must maintain SRs to exit routers

Further Overhead Reductions

Improve source selectivity.

- Initiate only for transport-layer flows
- Maintain SRs only for active destinations

Source could utilize transport-layer feedback.

- Don’t reissue RP if things are fine
- Reissue RP when (RTT + RP flow timer expires)

Policy Concerns

Illicit SR could violate usage policy.

Firewall rules for untrusted interface:

- Don’t reply to RPs
- Discard arriving OvIP packets

For OvIP flows allowed to cross firewall, the destination can perform IP header comparison.

Support for Quality-Of-Service can be provided.

- SR field contents privately defined
- Include QOS subfield --- such as priority

Prior Related Work

IP Switching

- Label swapping, distributed state
- Link-layer specific

Tag Switching

- Label swapping, distributed state
- Not link-layer specific

Sirpent

- Source routed, no distributed state
- Not link-layer specific
- Symmetric fwd,,bkwrd path so not Internet compatible

SUPPORTING SLIDES

SUPPORTING SLIDES

SUPPORTING SLIDES

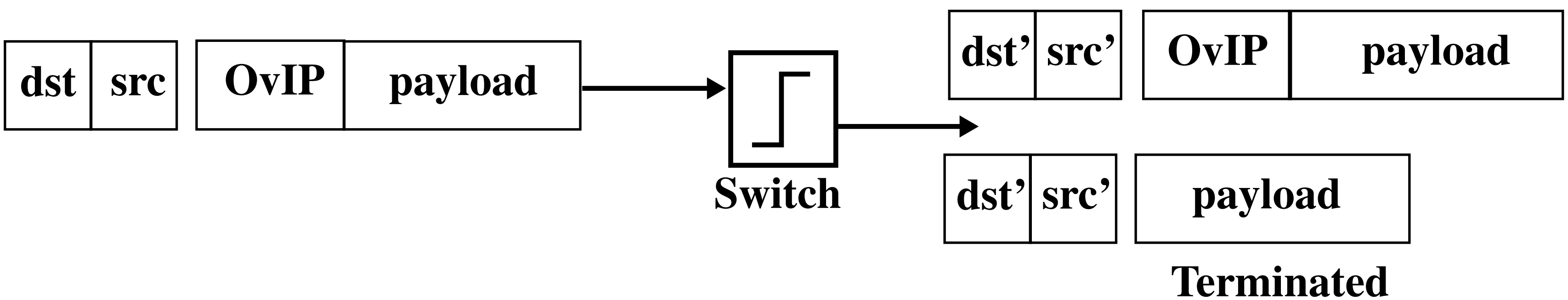
SUPPORTING SLIDES

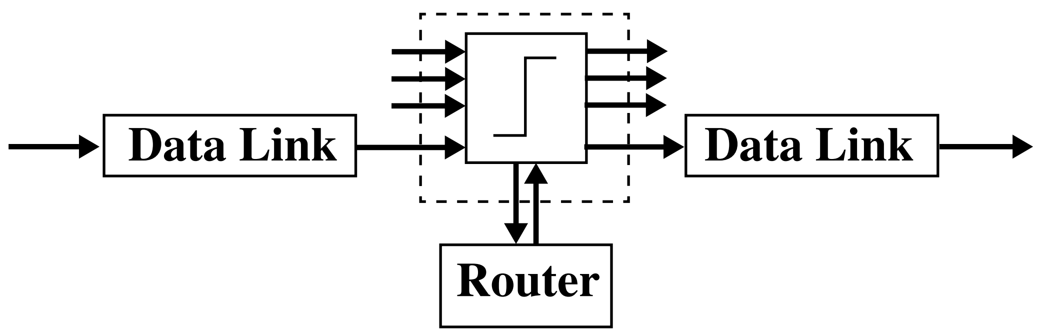
SUPPORTING SLIDES

OvIP Switch

Characterized by:

- Passive operation, stateless **(no forwarding tables)**
- Non-OvIP messages sent to router





OvIP Behavior

Need to determine:

1. Overhead
   -- RPs per switched packet **(compute and bandwidth)**
   -- Flow tables **(storage in initiator)**
2. Advantage **(% of traffic switched via OvIP)**
3. Criteria that min-max the above for OvIP

This can be investigated by examining IP packet traces.

OvIP Simulation

Adopt simple OvIP initiator model:

- Create OvIP source route per destination flow
- Refresh SRs every 10 sec., remove if idle for 10 sec.

Simulate using IP trace data.

- Overhead: #RPs issued / #packets switched
- Advantage: % traffic switched

Overhead	Advantage	Mean Flows/sec.	Flows
.025	**99.0**	**163**	**20618**

Trace: DEC-PKT-1, 2.15 Mpkt of TCP over 15 min.

Reducing OvIP Overhead

Potential exists for reducing overhead below 2.5%.
Don't create OvIP source route when . . .

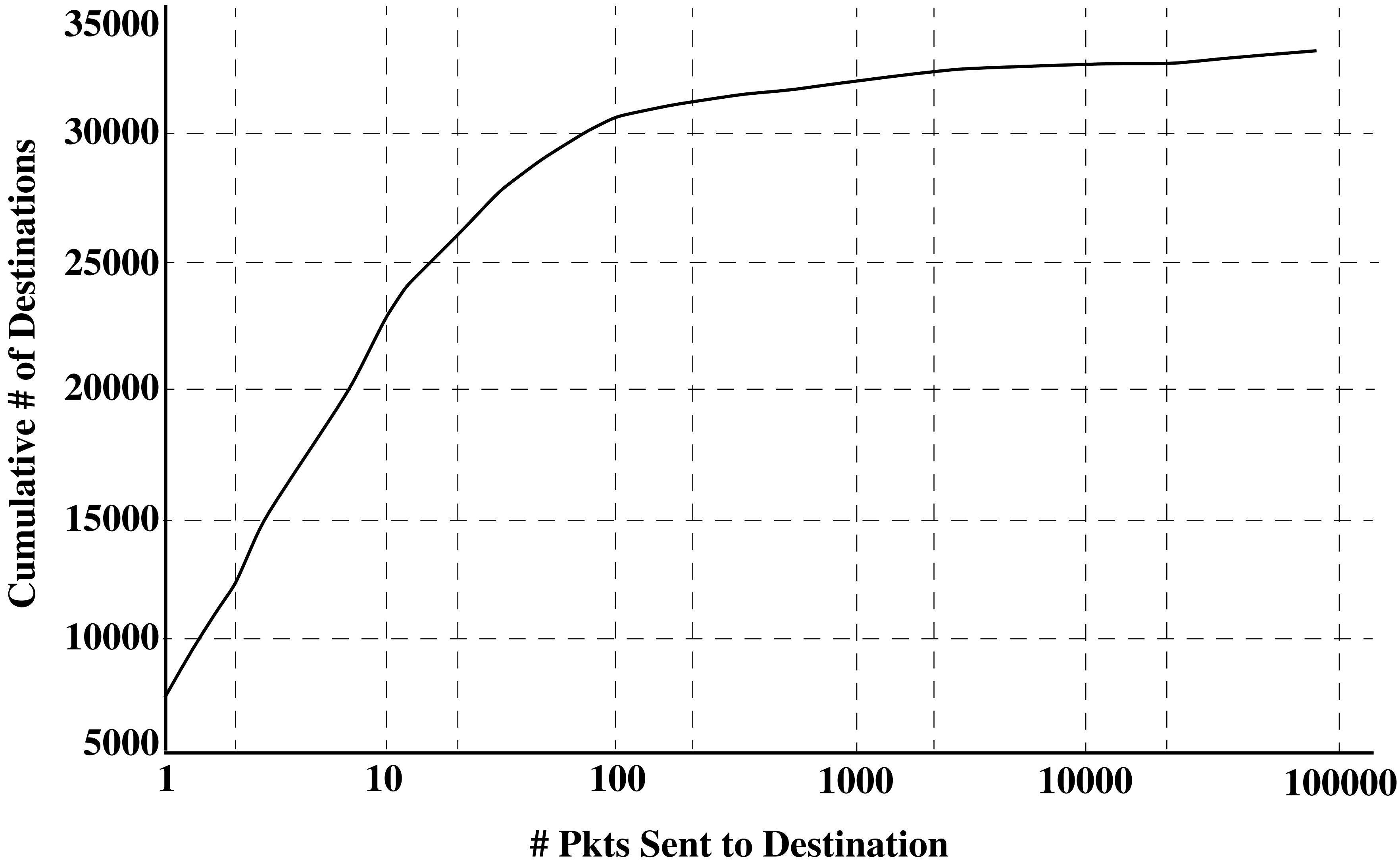
- Only few packets are sent across a flow
- Frequency of packets sent across flow is low

Are those reasonable criteria?

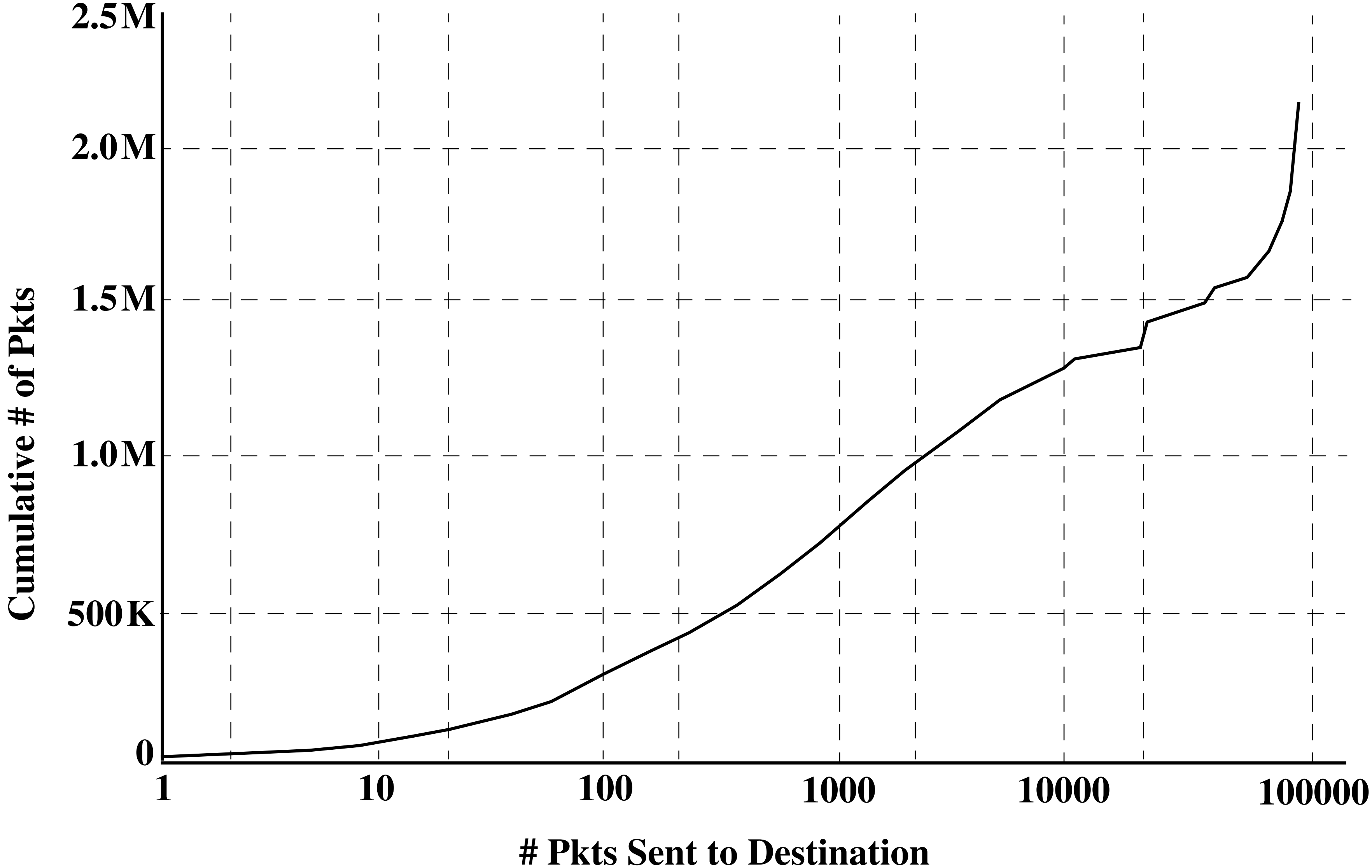
- What is distribution of #packets per destination flow?
- Is frequency of pkts sent per destination flow too low?
  (10 sec. refresh and 1% overhead ⇒ 10 pkts/sec.)

Will look at #pkts/flow first.

Fix-West Trace (2.2 Mpkts, 5 min., 2/22/97)



Fix-West Trace



Suggestion: Count-Based Threshold

Use OvIP only when flow has seen ≥ *n* packets.

Threshold	Overhead	Advantage	Mean Flows/sec.
1	.025	99.0	163
10	.017	95.2	106
15	.016	93.6	97
20	.015	92.4	92

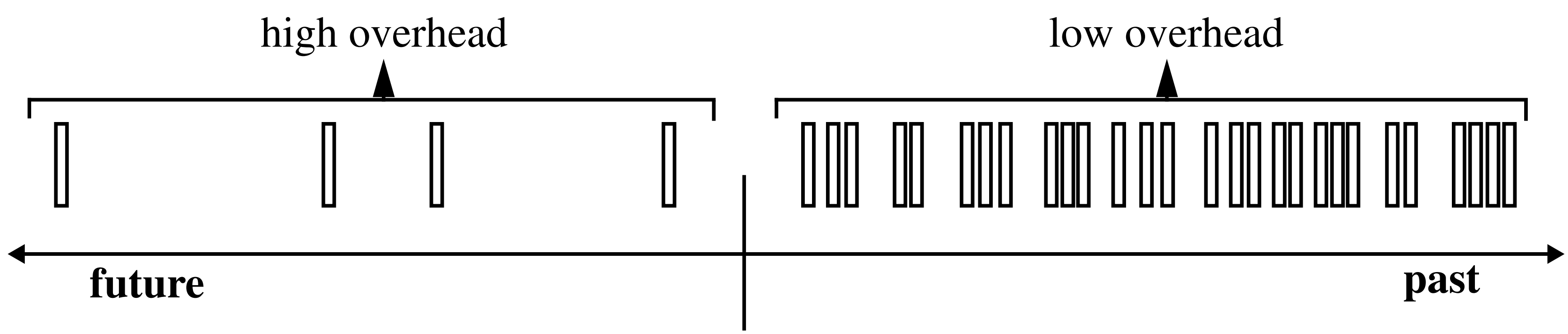
Trace is DEC-PKT-1, 2.15 Mpkt of TCP in 15 min.

Threshold of 20 reduces overhead by 40% at little cost.

Frequency-Based Threshold

Use OvIP when smoothed Inter-Arrival Time falls below limit.

<u>Flow</u>



Questions:

- Does frequency threshold perform better than a count?
- Should a low-pass or a notch filter be used?

SUPPORTING SLIDE

SUPPORTING SLIDE

SUPPORTING SLIDE

SUPPORTING SLIDE

SUPPORTING SLIDE

Drawback of Threshold-Based Overhead Reduction

RP overhead depends upon flow rate.

- Low-freq. flows produce excessive overhead
- 28.8 modem produces a low-freq flow

Observation:

Transport-layer protocols provide forward progress information via their reverse path.

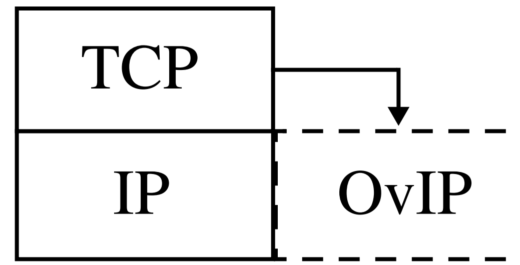
- To send RPs when flow is making progress is a waste
- Source retransmission indicates possible routing change

Transforms RP overhead from being . . .

frequency based $\Rightarrow$ route-quality based.

Further Overhead Reduction

Utilize transport-layer information.



- Don’t reissue RPs while connection is making progress
- RP overhead could approach one/flow

Trace suggests bound on overhead ~ 0.003 RPs / data packet.

Summary: OvIP ---Plus and Minus

Pluses

- Simple: No flow state distribution
- Stateless: Passive operation, attack resistant
- No table limitations forcing flow aggregation
- End-to-end compatible **(not dependent on particular link-layer)**

Minuses

- Increased packet length
- Route-Ping overhead
- Fundamentally unicast