Towards Differentiated Services for the Internet

Van Jacobson
(van@ee.lbl.gov)
Network Research Group
Berkeley National Laboratory
Berkeley, CA 94720

Bay Networks Architecture Lab
Boston, MA

November 25, 1997

What problem are we solving?

Give “better” service to some (at the expense of giving worse service to others — despite QoS fantasies to the contrary, it’s a zero-sum game).

Why?

- Some applications (voice and video) think they need it
- Some users (e.g., ISPs) need better traffic control
- Mission-critical applications will stay on leased lines until this exists
- Economics says it's the only long-term way to make a multi-provider, commercial Internet self-sustaining.

Who controls what's special — users or institutions?

If users, then solution is trivial – users individually decide which of their traffic is most important and attach labels to convey this to the network. There are no trust or coordination issues.

Unfortunately, most of the problems and essentially all the demand call for institutional control of sharing.

Since users can't get whatever they want, there's incentive to steal and architecture must include good security.

(This is especially important if design attempts to limit state in the network.)

Is the control end-to-end, hop-by-hop, intra-domain, inter-domain, per-path, or per-boundary?

Yes.

(The Internet is big, there must be a win to all forms of incremental deployment or the service will never get deployed.)

What is the service: “Better best effort” or “Virtual leased line”?

Yes.

(Current demand for the former seems to be mostly intra-domain traffic control while the latter seems to be inter-domain service offerings. But there is demand for both.)

What are the target applications / protocols?

Bad question. In 1978, the answer was RJE. In 1988, email/ftp. In 1998, probably web. This too will change.

IP/TCP/UDP/IGMP/OSPF/BGP work for any application.
Differentiated services must too.

Design Constraints — Scaling

A Differentiated Service mechanism must work at the scale of the Internet (e.g., millions of networks) and at the full range of speeds of the Internet (e.g., Gb/s links). To get that kind of scaling the design must

- push all the state to the edges, and
- force all per-conversation work (e.g., shaping, policing) to the edges.

Design Constraints — Scaling

$\Rightarrow$ Edge-only state suggests that special/normal service indication must be carried in the packet.

$\Rightarrow$ Administrative diversity and high speed forwarding both argue for very simple semantics on that indication. E.g., one or two bits of special/normal.

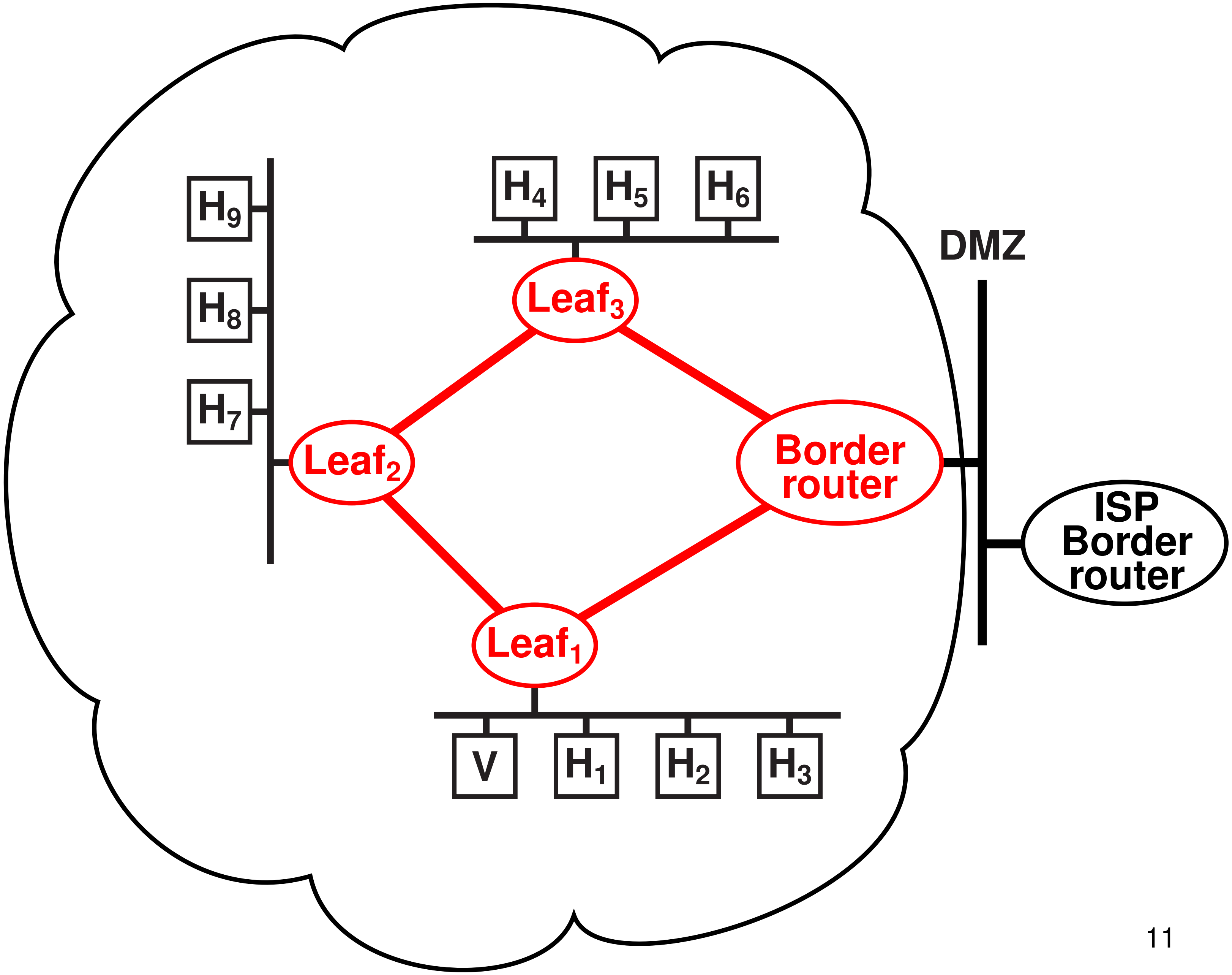
$\Rightarrow$ No state in center means everything but edge will see only aggregates (potential fairness problems).

Dave Clark (MIT LCS) has proposed “**edge-tagging**” as a scalable way of offering differentiated services.

- Leaf router adjacent to the source(s) has traffic signature for “special” traffic and “profile meter” giving its characteristics.

- That router “marks” (sets IP precedence field) in all special traffic that conforms to profile meter.

- All routers unmark all other traffic.


H9
H8
H7
H4
H5
H6
Leaf3
Leaf2
DMZ
Border router
ISP Border router
Leaf1
V
H1
H2
H3

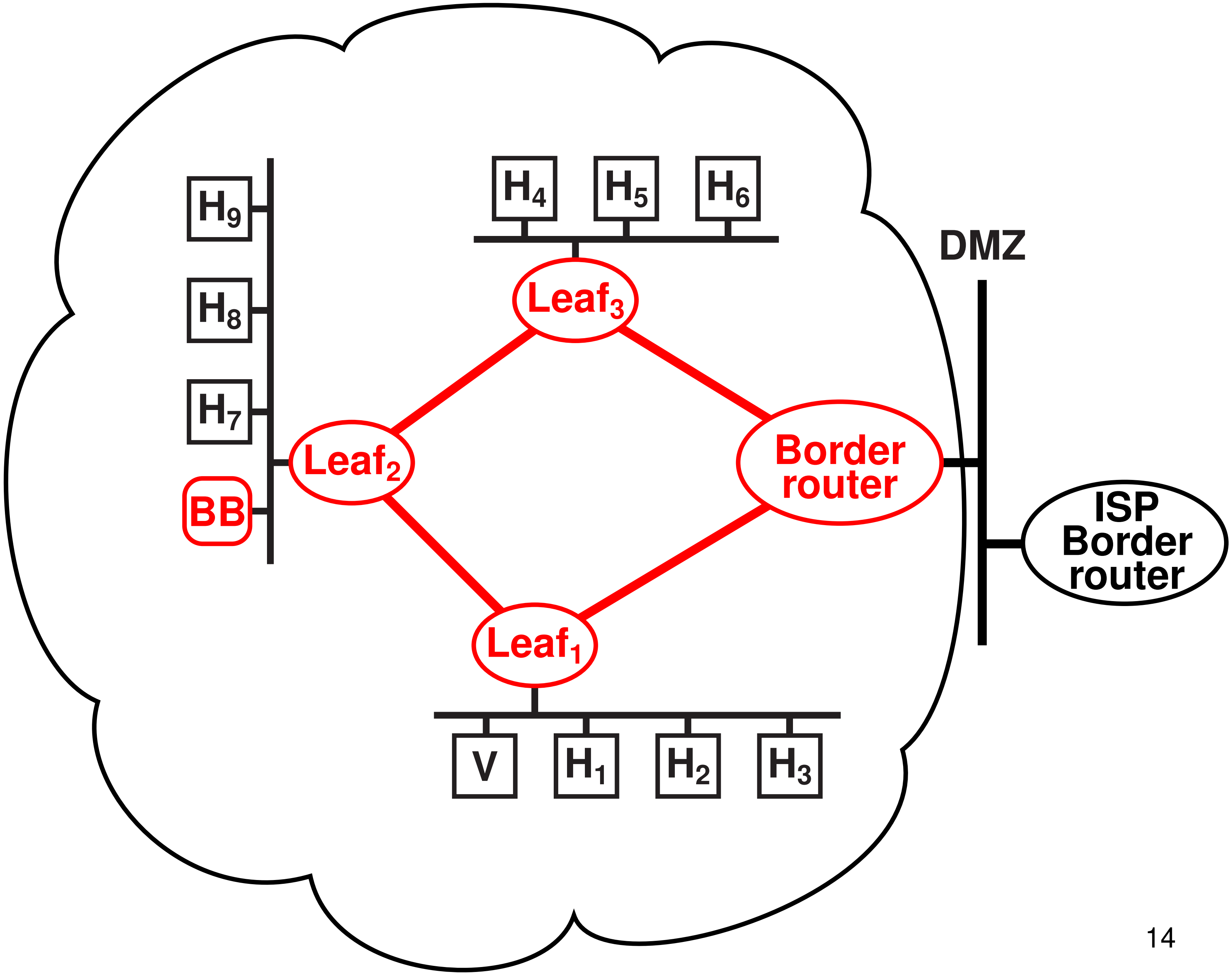
But there are still problems:

- Who decides what users get to request special service?
- Where is organizational policy on use of limited bandwidth implemented?
- Who tells the edge router what to tag?
- Who makes sure that simultaneous uses of special service fit within allocation?

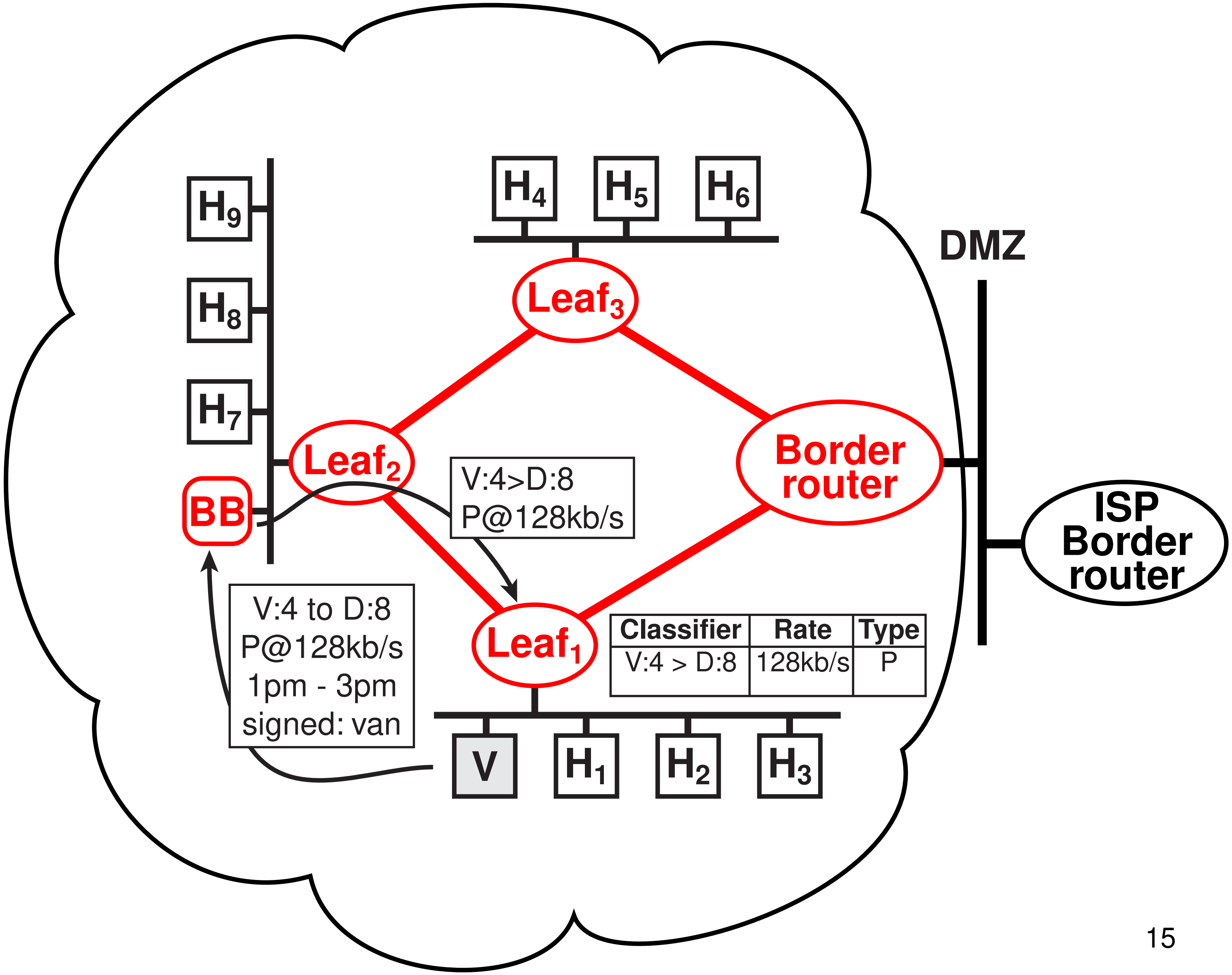
Answer: Introduce a **Bandwidth Broker** (BB) to be repository of policy database of priority and limits for user & project access to special bandwidth. Repository includes user credentials so requests can be authenticated.

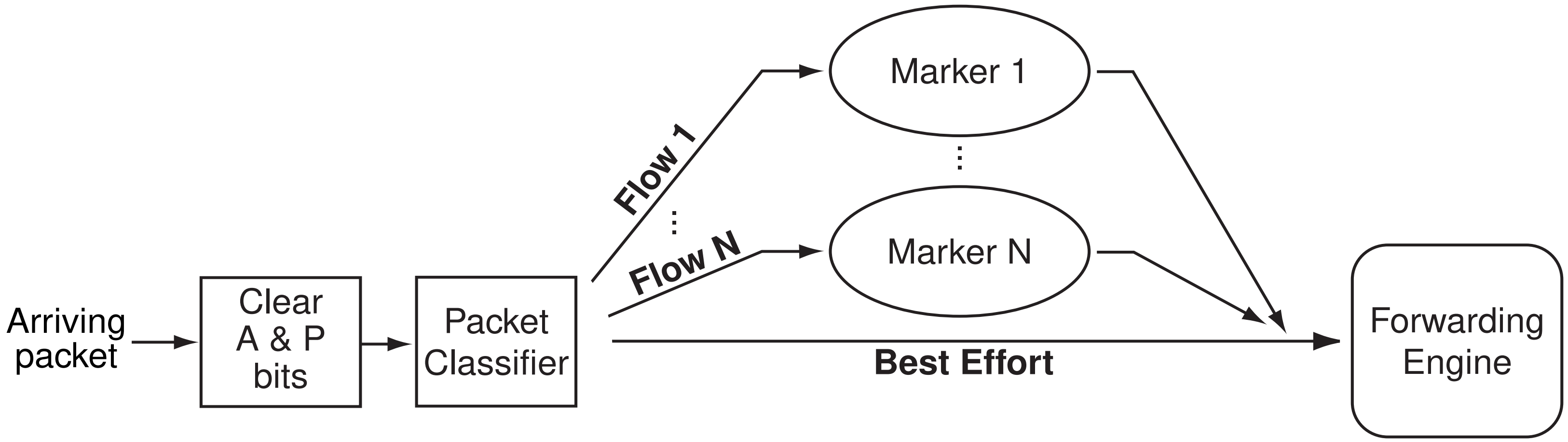
BB is part of network infrastructure so can have trusted, secure association with all routers.

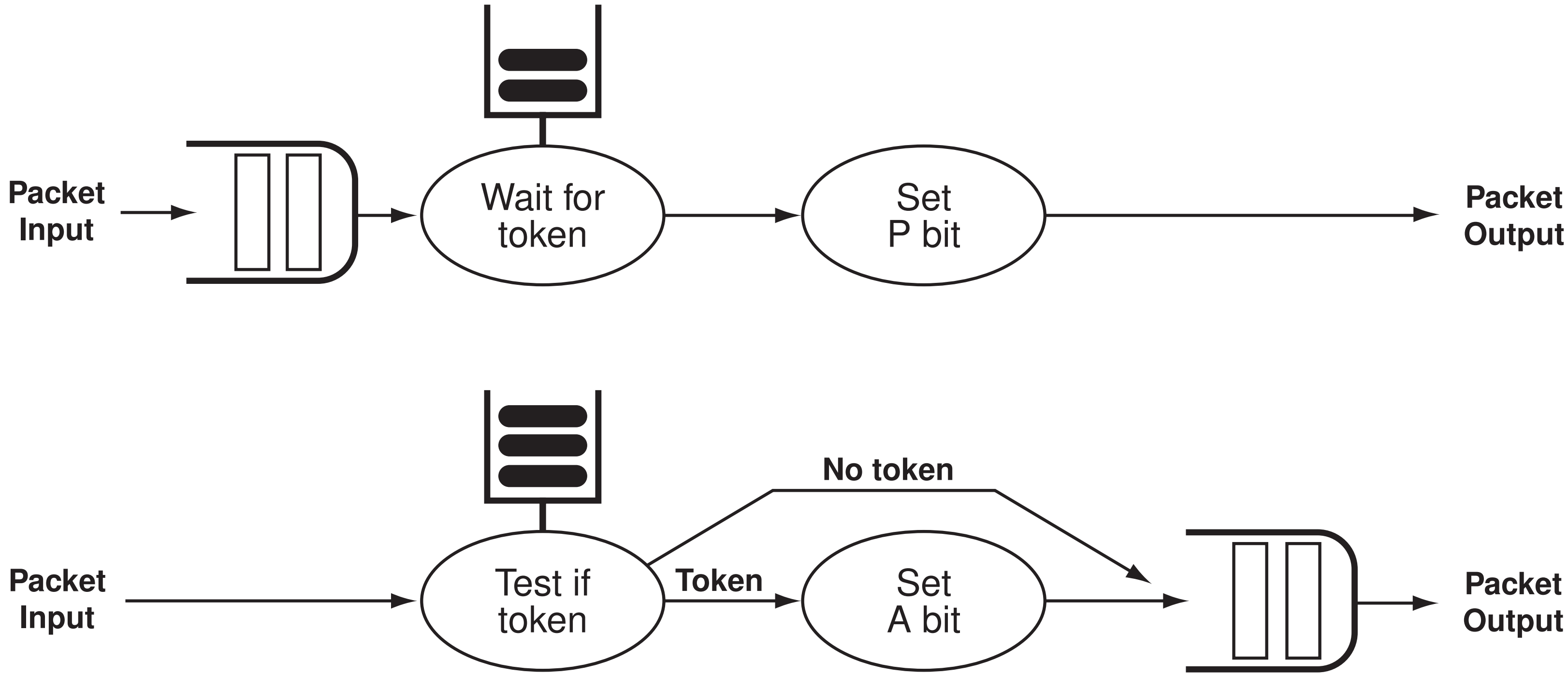
Requests go from user to BB (so it can record use and resolve conflicts) then to appropriate router so security model is well-founded.



H9
H8
H7
BB
H4
H5
H6
Leaf3
Leaf2
Leaf1
Border router
DMZ
ISP Border router
V
H1
H2
H3


H9
H8
H7
BB
H4
H5
H6
DMZ
Leaf3
Leaf2
Border router
ISP Border router
V:4>D:8
P@128kb/s
V:4 to D:8
P@128kb/s
1pm - 3pm
signed: van
Leaf1
Classifier | Rate | Type
V:4 > D:8 | 128kb/s | P
V
H1
H2
H3


Arriving packet
Clear A & P bits
Packet Classifier
Flow 1
Flow N
Marker 1
Marker N
Best Effort
Forwarding Engine


Packet
Input
Wait for
token
Set
P bit
Packet
Output
Packet
Input
Test if
token
No token
Token
Set
A bit
Packet
Output

Is this sender or receiver based?

The question is broken.

Special service allocation follows administrative hierarchy, not topological one.

Real question is whether design allows something from outside the sender’s domain to frob its Bandwidth Broker (it does).

If so, sender, receiver or kindly third party can all make sender’s traffic become special.

Design Constraints — Interdomain service

Almost all Internet traffic crosses *many* administrative boundaries. End-to-end service implies that all those independent units agree to treat the traffic as special. Multilateral agreements rarely work.

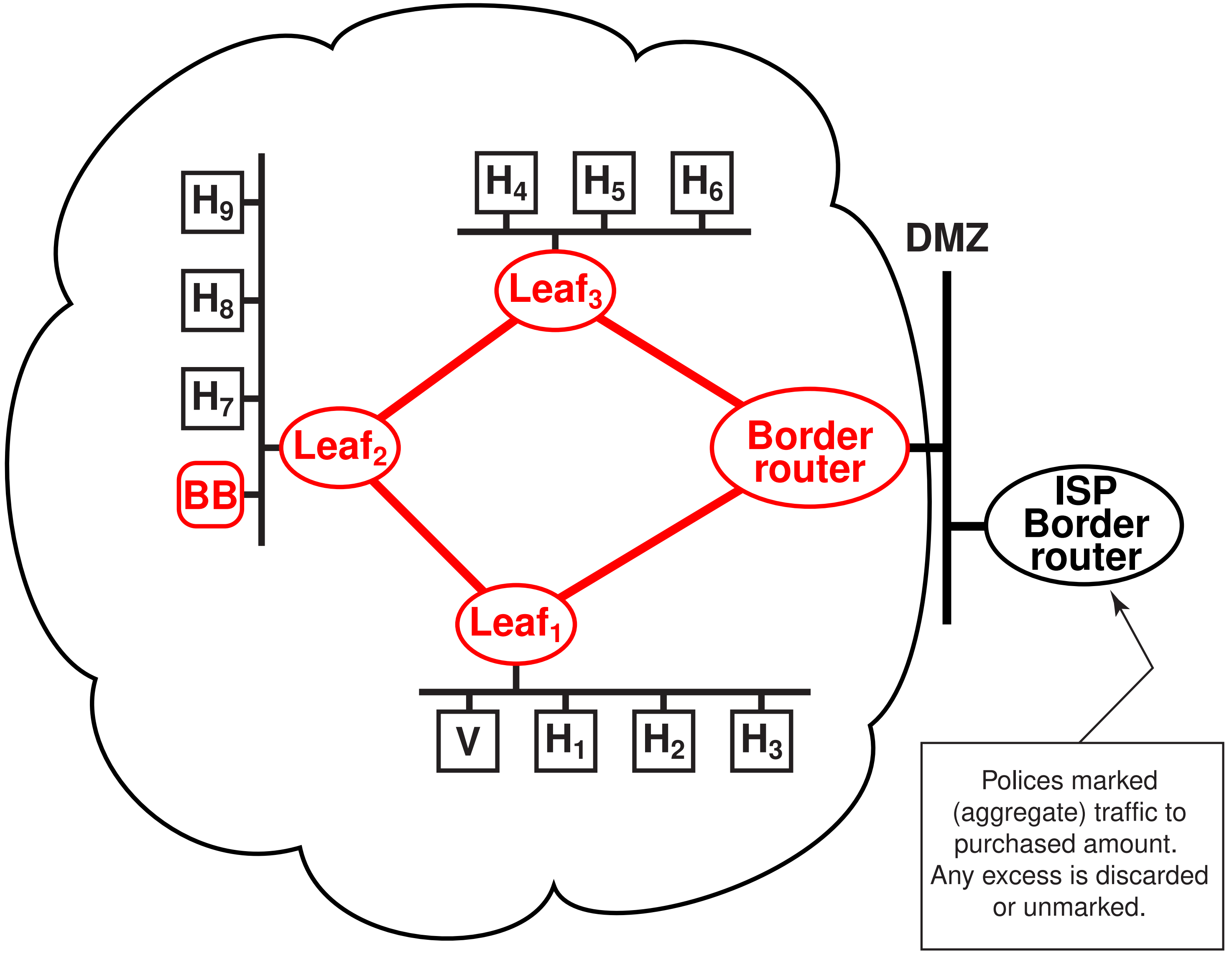
ISPs are competitive enterprises. They act in their own best interests and, where necessary, against the interest of their competitors.

Design Constraints — Interdomain service (cont.)

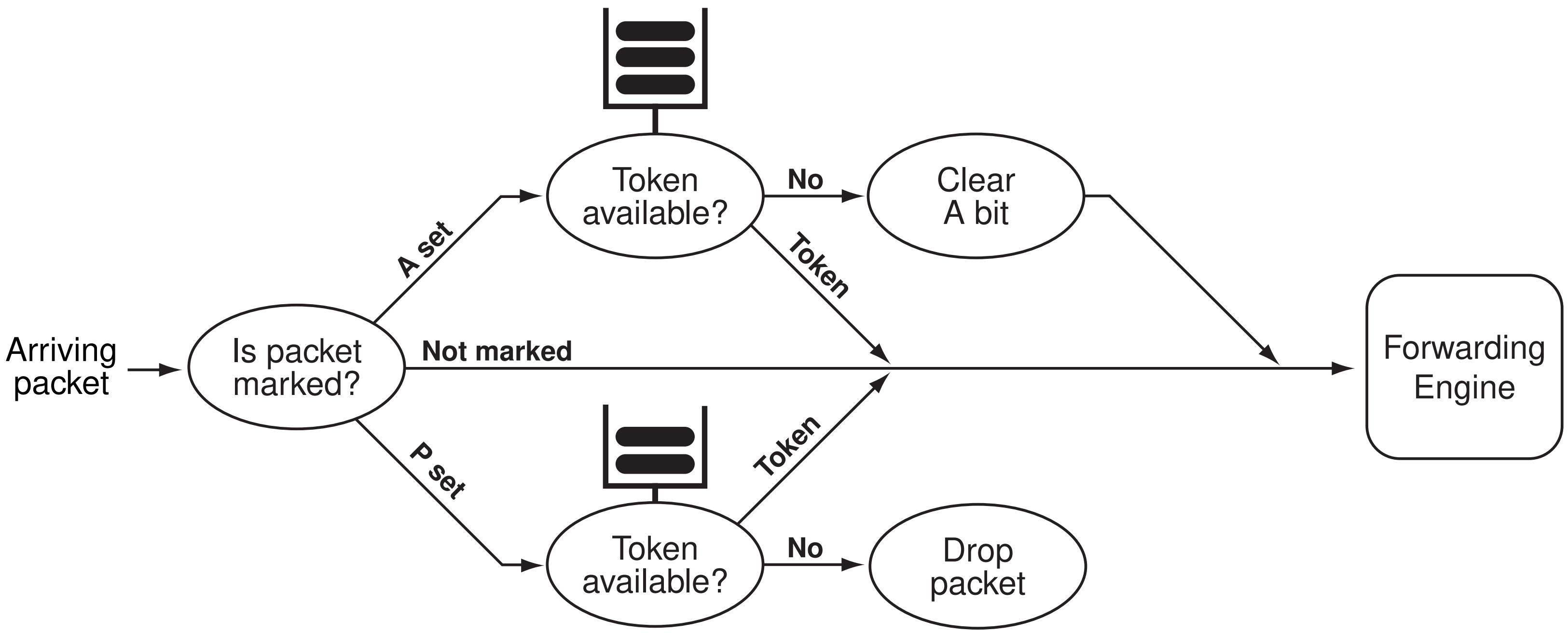
⇒ End-to-end service should be constructed from bilateral agreements.

⇒ Service must not require extending trust or control across administrative boundaries. (Upstream can't force extra work on downstream.)

⇒ Must have fault isolation (a customer shouldn't be able to trash another but is welcome to trash himself).



H9
H8
H7
BB
H4
H5
H6
Leaf3
Leaf2
Leaf1
Border router
DMZ
ISP Border router
V
H1
H2
H3
Polices marked (aggregate) traffic to purchased amount. Any excess is discarded or unmarked.



Arriving packet
Is packet marked?
A set
Not marked
P set
Token available?
No
Clear A bit
Token
Token
Token available?
No
Drop packet
Forwarding Engine

Design Constraints — Service characterization

For economically viable service, customers have to know what they're buying.

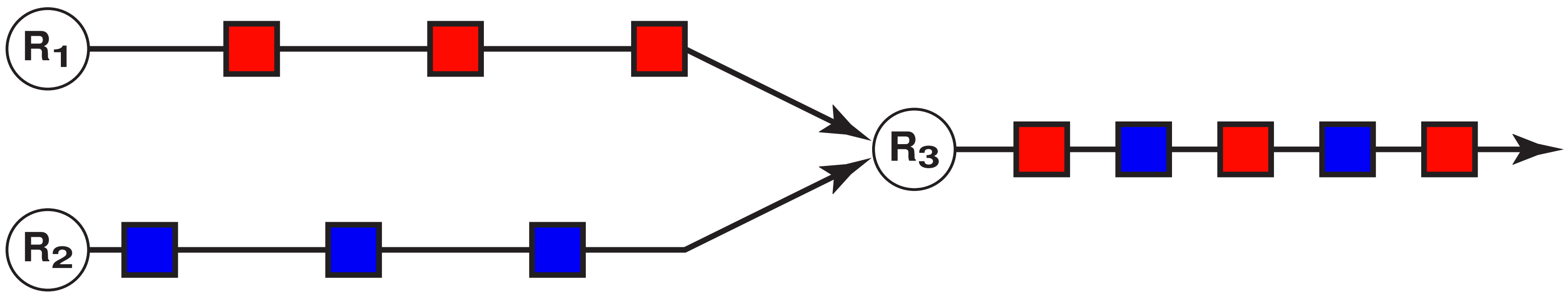
- Delivered service *can't* depend on other people's traffic.

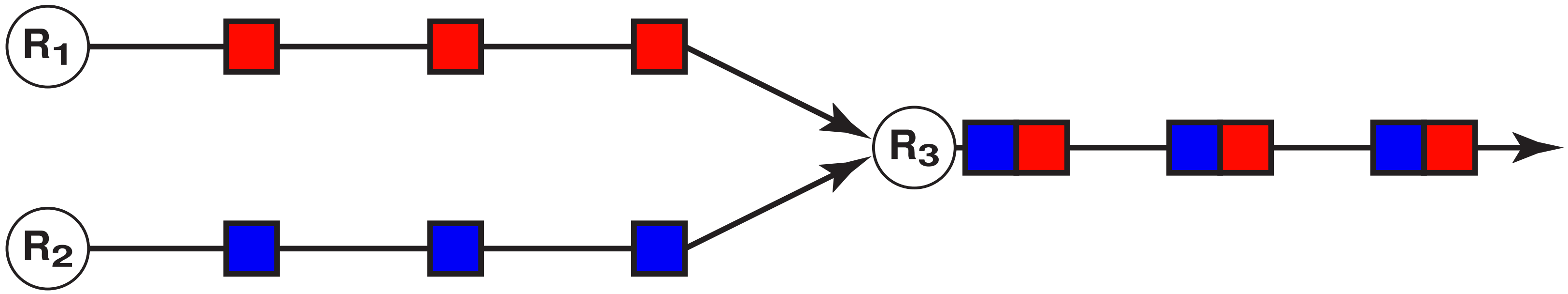
- The customer *must* be able to measure conformance.

$\Rightarrow$ ISP must not "over sell" special service (simple queuing theory: if aggregate inflow $>$ outflow, delivered bandwidth, drop rate and/or delay can be *arbitrarily bad*).

$\Rightarrow$ Even if service is limited to capacity, aggregation can cause some users to get poor or non-existant service unless **traffic shaping** done at borders.

Even for fixed rate traffic, *phase differences* create bursts in an aggregate:





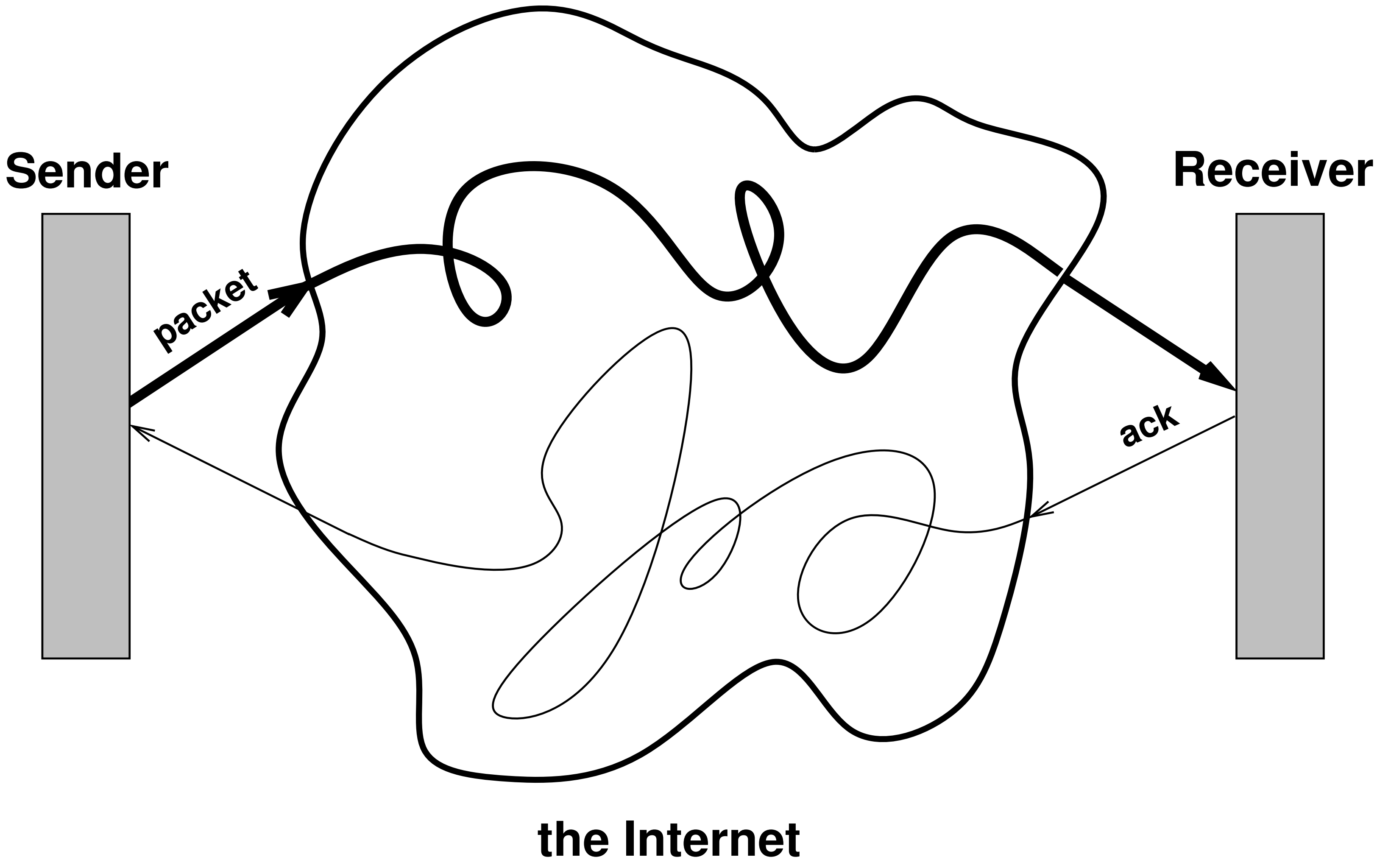
- If no shaping is done, burst size scales like 2^N for source N hops away.

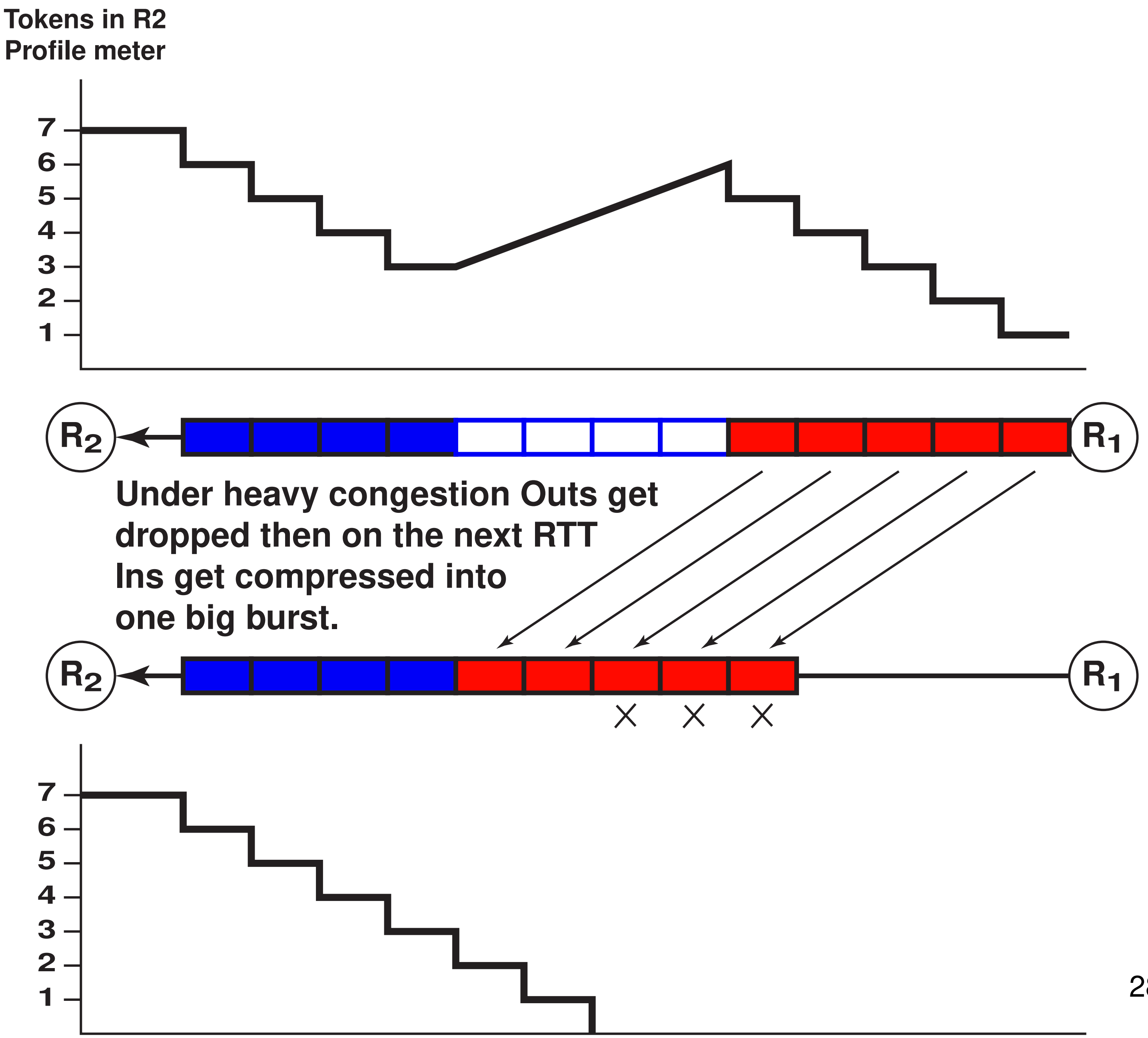
- If shaping is done at borders, burst size scales like average in-degree in region.

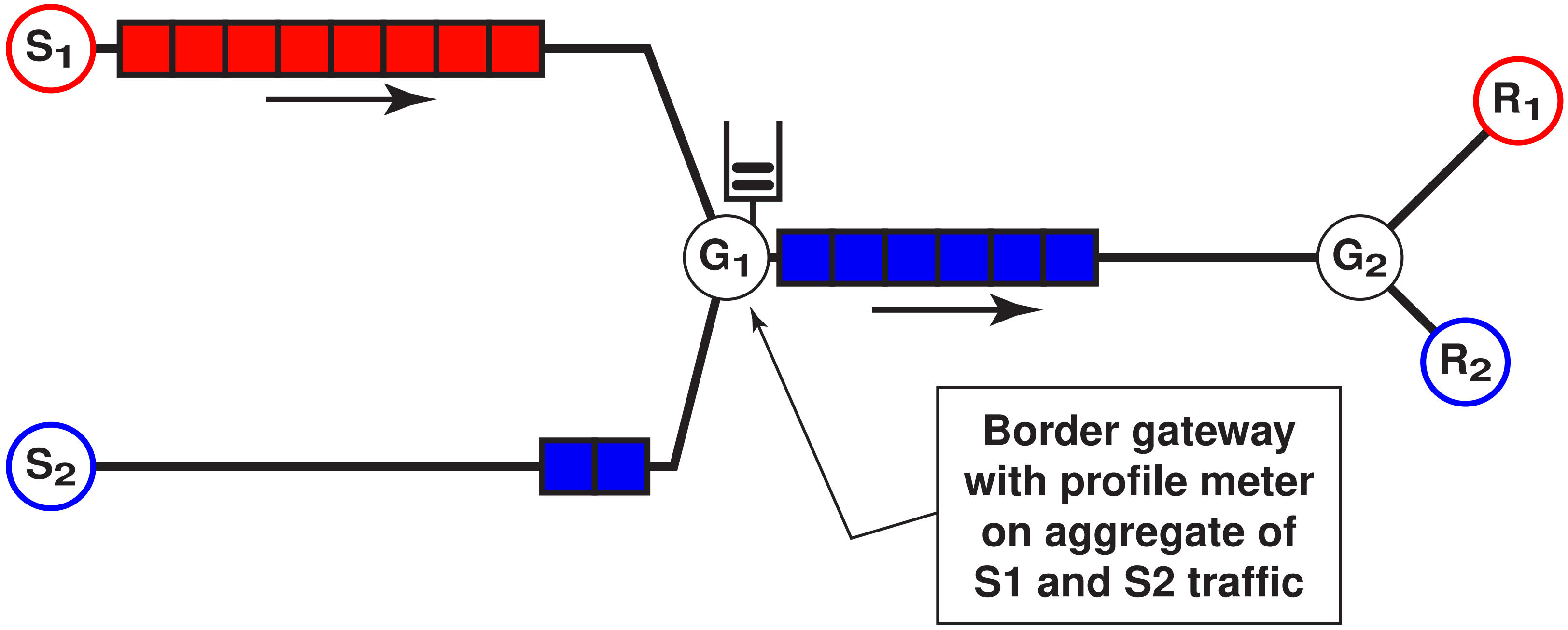
With no shaping, the system is very brittle (special users may see very poor service depending on their location in topology and/or competing traffic.

Even ignoring phase interactions, drop-preference schemes can result in extreme unfairness due to interaction with the traffic structure induced by transport protocol dynamics.

This is a more ubiquitous problem but it could also be remedied if the service model allowed shaping at upstream border routers.


Sender
Receiver
packet
ack
the Internet


Tokens in R2
Profile meter
7
6
5
4
3
2
1
R2
R1
Under heavy congestion Outs get
dropped then on the next RTT
Ins get compressed into
one big burst.
R2
R1
7
6
5
4
3
2
1



S_1
S_2
G_1
G_2
R_1
R_2
Border gateway
with profile meter
on aggregate of
S1 and S2 traffic

Setup transaction can be a major part of operating costs. There are two design extremes:

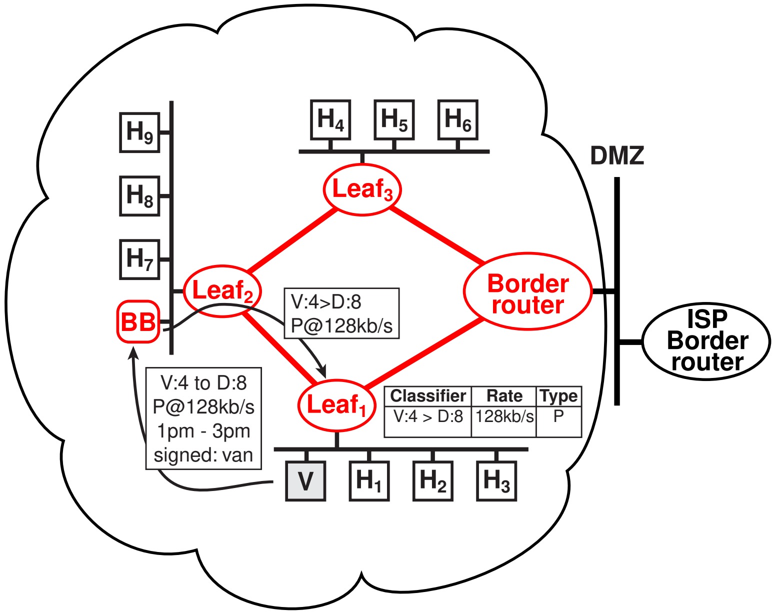
Everything dynamic $\longleftrightarrow$ **Everything pre-negotiated**

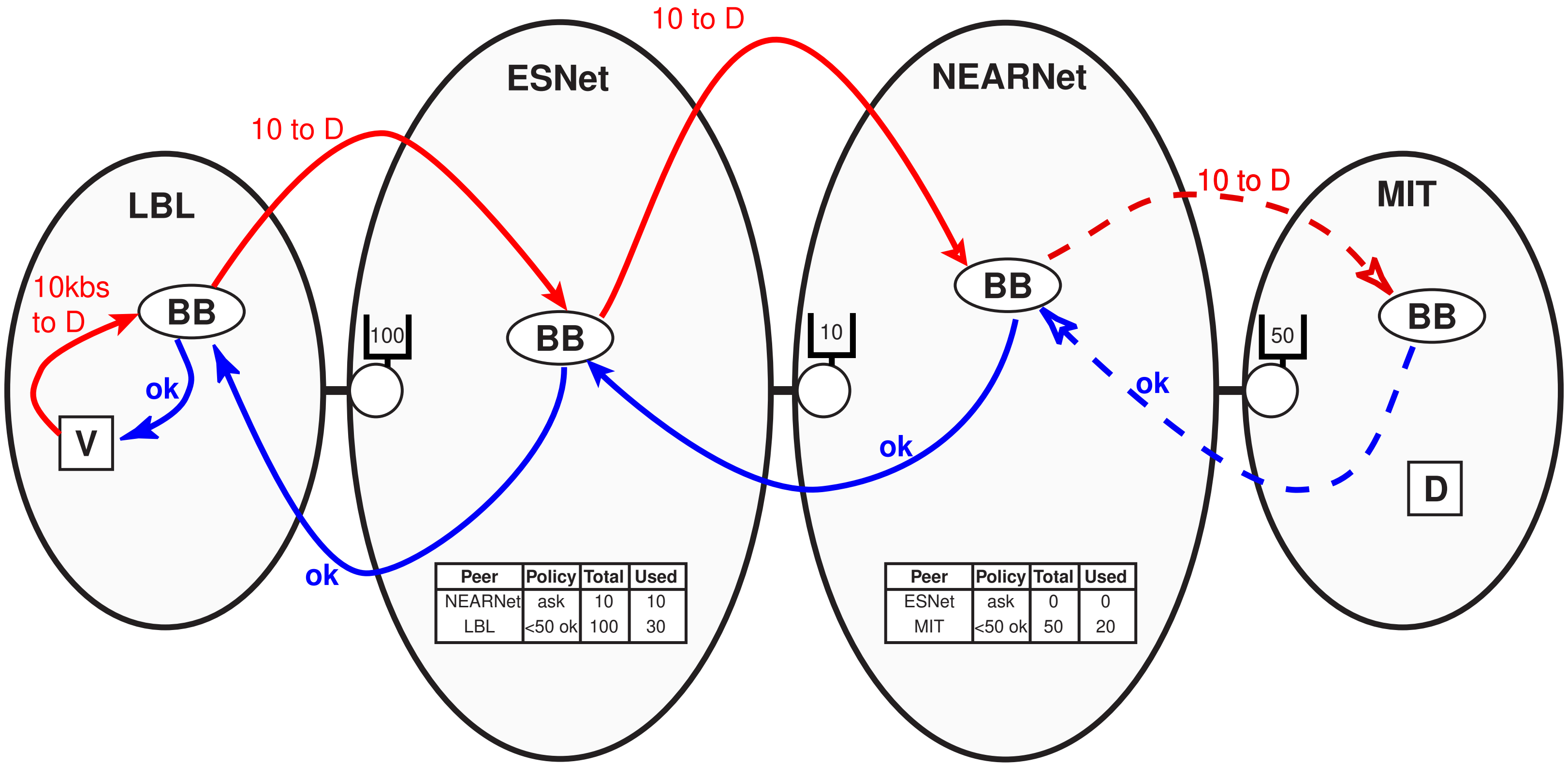
The left side has good resource efficiency, high transaction costs, and poor cost control. The right side is the opposite. Neither works for everyone and the design should let you to operate at any point on the line.

Note that aggregate demand can always be split into two components: predictable and exceptions. To tune transaction costs:

- pre-negotiate (and pre-purchase) predictable component,
- handle exceptions via on-demand purchase on “spot market”,
- do efficient “internal” allocations of pre-negotiated service,
- make customer/provider agreement include hooks for setup transaction propagation.

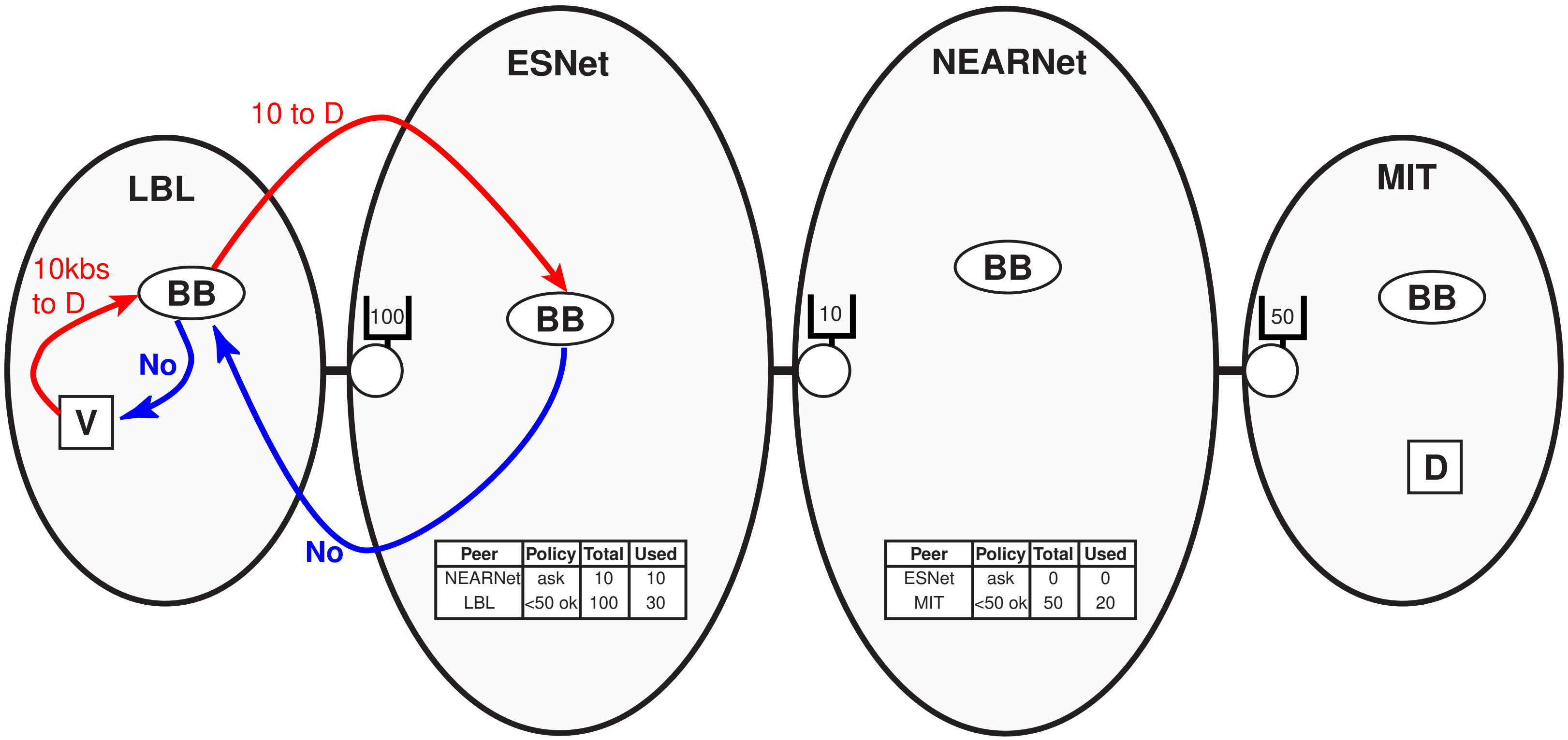
The Bandwidth Broker, necessary to incorporate institutional policy into the bandwidth sharing decisions, provides all the machinery needed to handle the no-signaling (pre-negotiated) case:





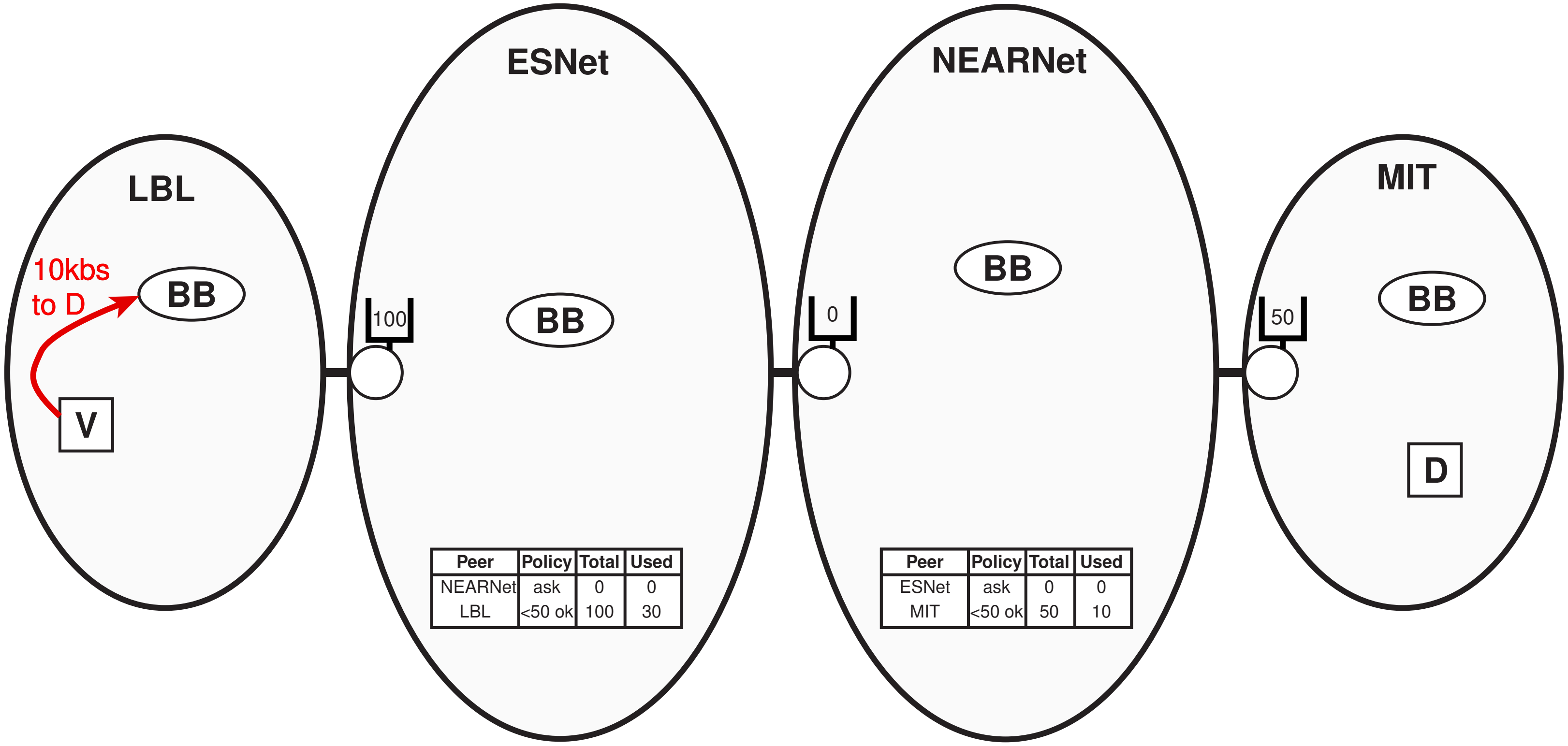
Peer	Policy	Total	Used
NEARNet	ask	10	10
LBL	<50 ok	100	30

Peer	Policy	Total	Used
ESNet	ask	0	0
MIT	<50 ok	50	20



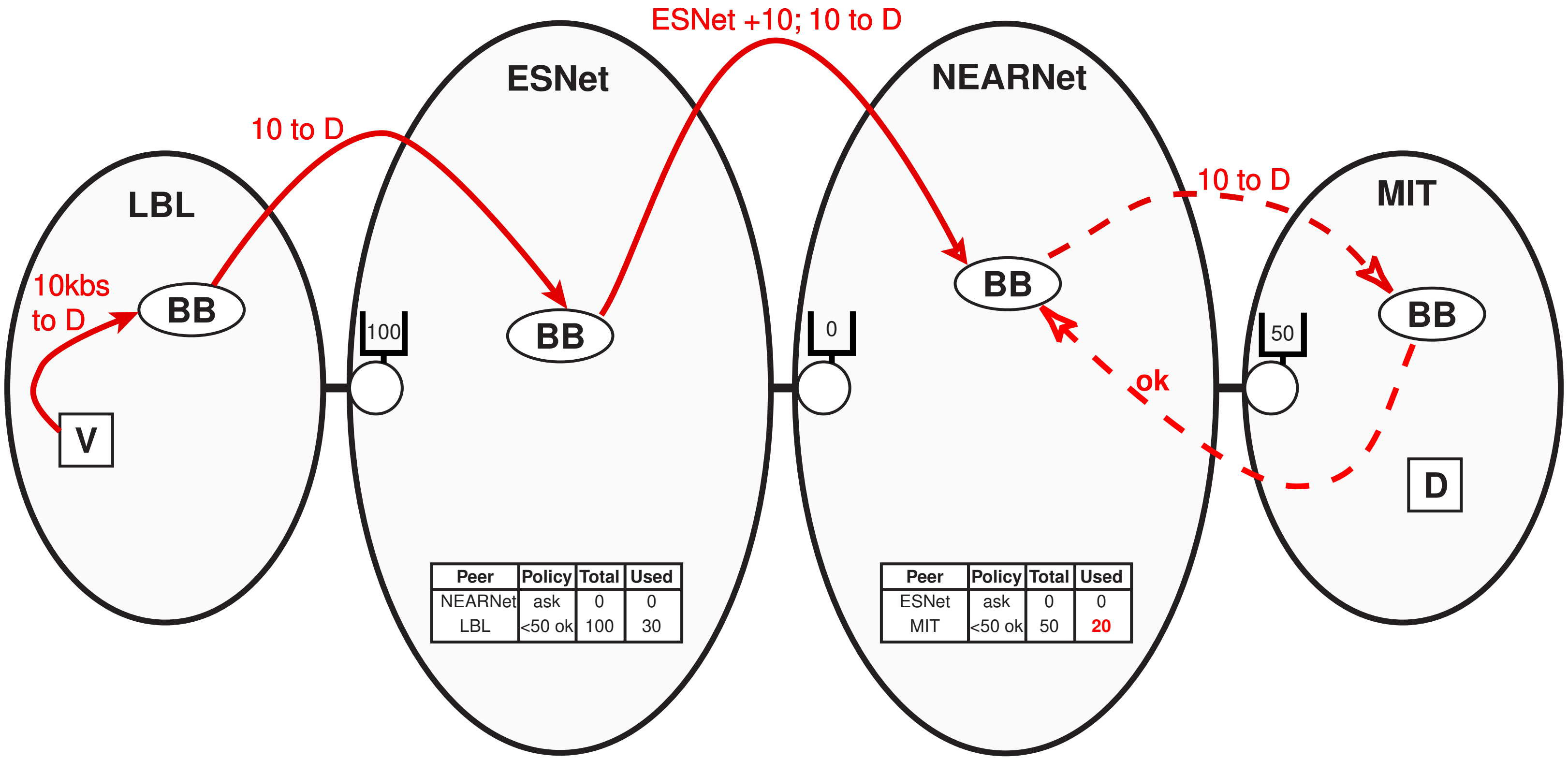
Peer	Policy	Total	Used
NEARNet	ask	10	10
LBL	<50 ok	100	30

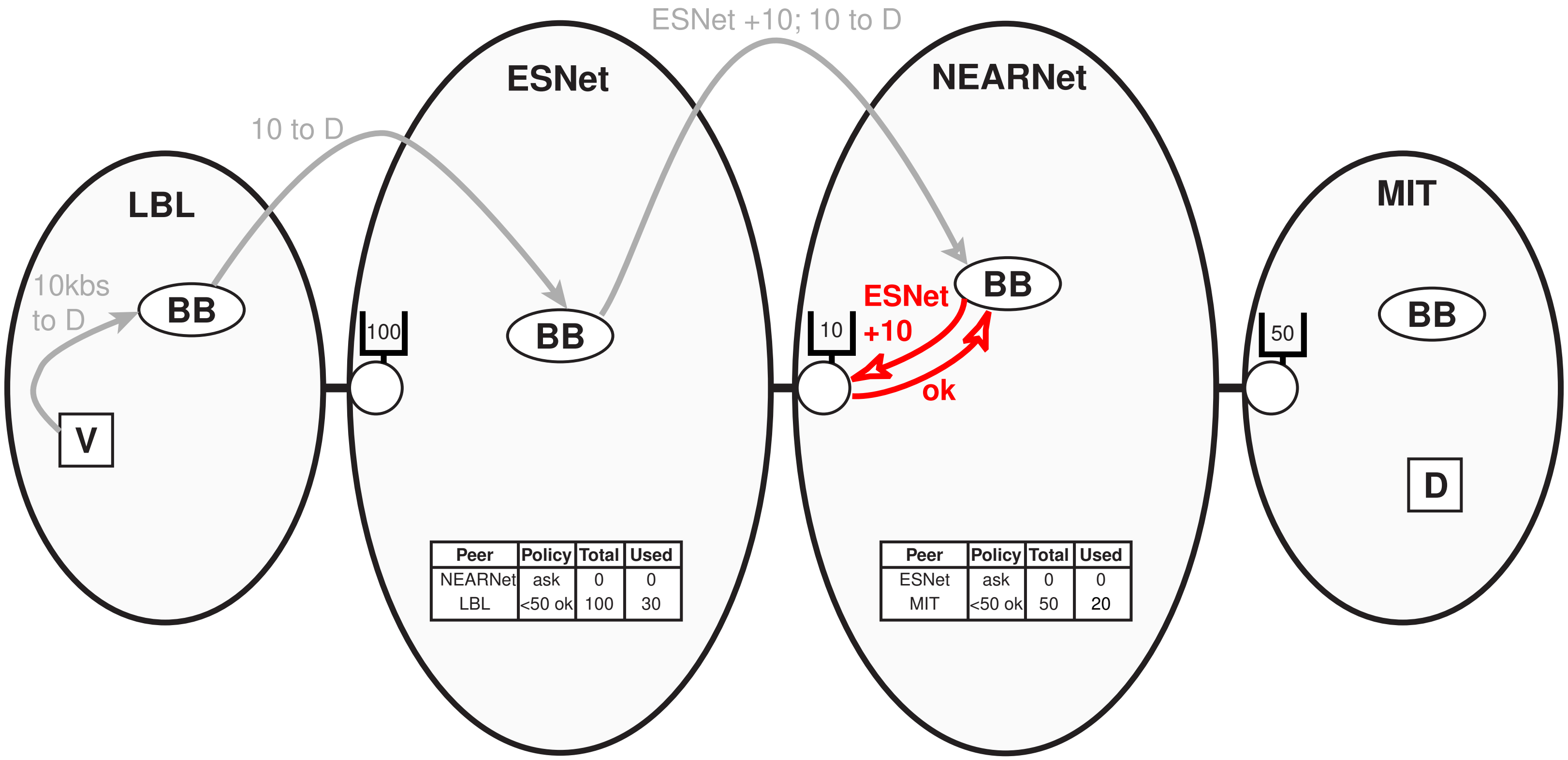
Peer	Policy	Total	Used
ESNet	ask	0	0
MIT	<50 ok	50	20

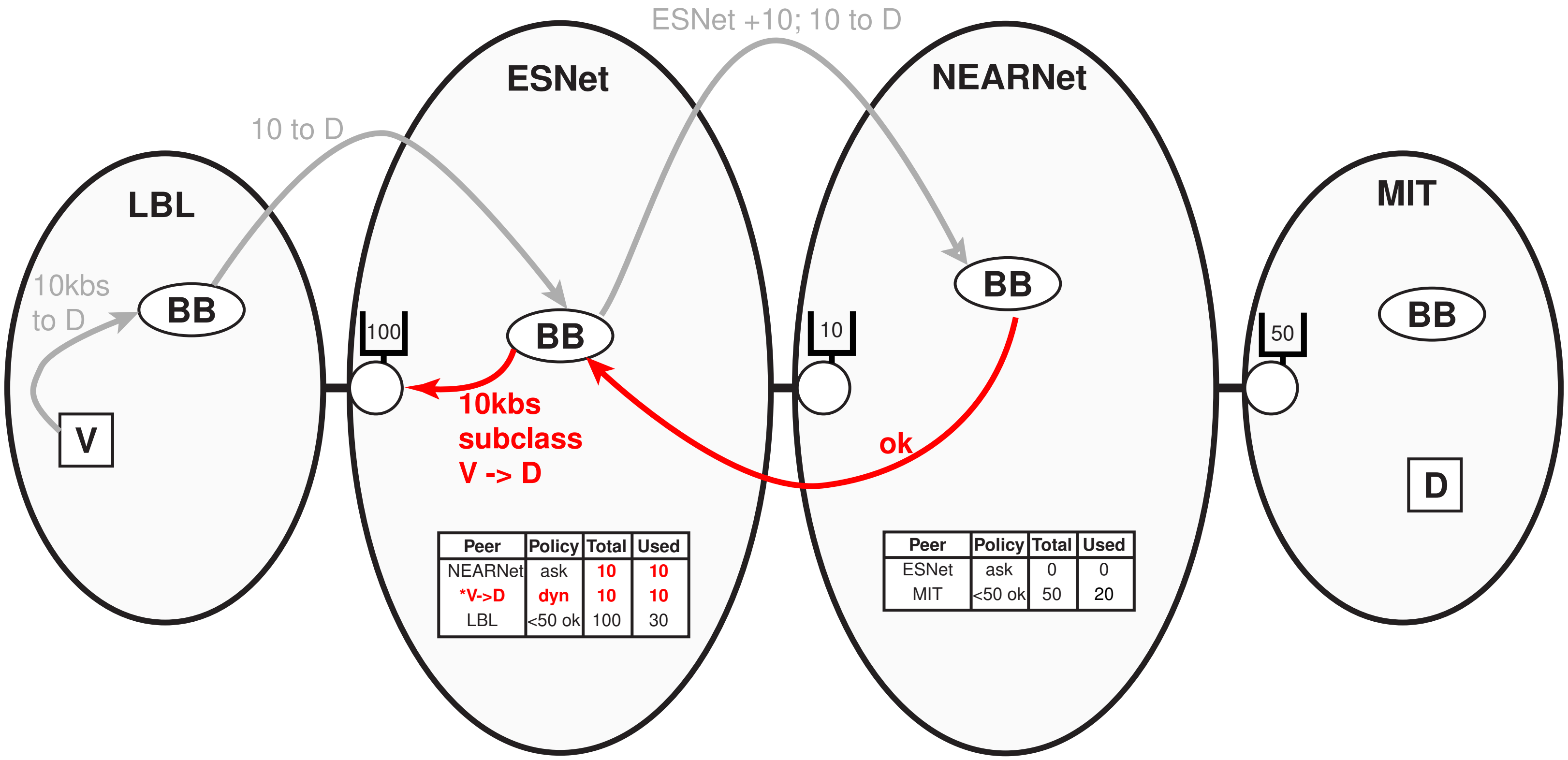


Peer	Policy	Total	Used
NEARNet	ask	0	0
LBL	<50 ok	100	30

Peer	Policy	Total	Used
ESNet	ask	0	0
MIT	<50 ok	50	10


ESNet +10; 10 to D
ESNet
NEARNet
10 to D
LBL
10 to D
MIT
10kbs
to D
BB
BB
BB
BB
100
0
50
ok
V
D
Peer | Policy | Total | Used
NEARNet | ask | 0 | 0
LBL | <50 ok | 100 | 30
Peer | Policy | Total | Used
ESNet | ask | 0 | 0
MIT | <50 ok | 50 | 20


ESNet +10; 10 to D
10 to D
LBL
10kbs
to D
BB
V
100
ESNet
BB
Peer | Policy | Total | Used
NEARNet | ask | 0 | 0
LBL | <50 ok | 100 | 30
NEARNet
BB
ESNet
+10
ok
10
Peer | Policy | Total | Used
ESNet | ask | 0 | 0
MIT | <50 ok | 50 | 20
MIT
BB
50
D


ESNet +10; 10 to D
10 to D
LBL
ESNet
NEARNet
MIT
10kbs
to D
BB
BB
BB
BB
V
D
100
10
50
10kbs
subclass
V -> D
ok
Peer | Policy | Total | Used
NEARNet | ask | 10 | 10
*V->D | dyn | 10 | 10
LBL | <50 ok | 100 | 30
Peer | Policy | Total | Used
ESNet | ask | 0 | 0
MIT | <50 ok | 50 | 20