VEGVÍSIR
systems

Easy and scalable Traffic Engineering with Traffic Dictator

Version 1.0
09.06.2024

Overview

Traffic Engineering is a set of techniques to steer certain types of traffic via a path different from the IGP-calculated shortest path. The reasons to deploy Traffic Engineering can include:

- Send some delay-sensitive traffic over a lower-delay path
- Reserve bandwidth for services
- Balance traffic across all available paths to avoid overloading some links while other links can be underutilized
- If an application has primary and backup data channels, traffic engineering can ensure they are forwarded over paths that don't use the same links

Another flavour of Traffic Engineering is Egress Peer Engineering (EPE) whereby an ingress router can choose not only the egress router but also egress peer where traffic will be forwarded.

Legacy Traffic Engineering

Since early 2000s, network operators have been deploying MPLS Traffic Engineering (MPLS-TE) with RSVP. By now, this is a mature and well-known technology but it has drawbacks:

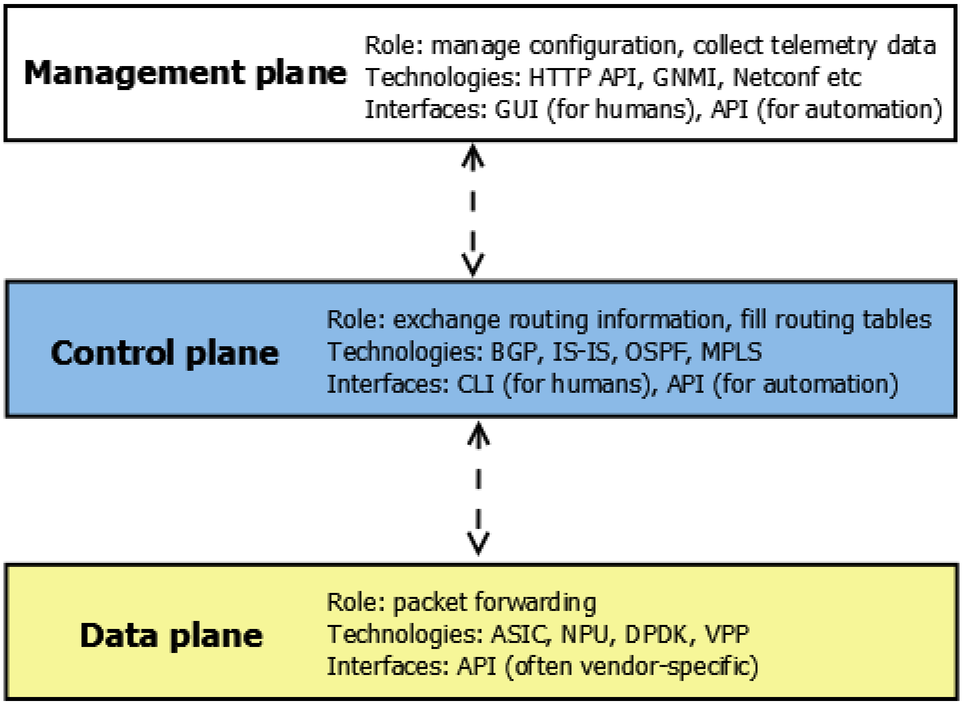
- Poor scalability
- Operational complexity
- Lack of support for ECMP and Anycast routing
- A very large set of features must be supported by all routers in the network which makes entry barrier very high and results in vendor lock-in
- Provisioning new services often requires manual configuration on many routers

Segment Routing

Segment Routing (SR) is a Software Defined Networking (SDN) technology which finds a perfect balance between traditional networking and SDN. It relies on standard protocols such as IS-IS, OSPF and BGP to provide basic network connectivity. Then an optional controller can be deployed to enhance the network with Traffic Engineering capabilities, provide bandwidth reservations and Egress Peer Engineering if required. Open standard protocols are used for router<>controller communication.

SR-TE controllers

Functionality of network devices is best understood through the distinction into Data, Control and Management planes. The picture below illustrates these roles:


Management plane
Role: manage configuration, collect telemetry data
Technologies: HTTP API, GNMI, Netconf etc
Interfaces: GUI (for humans), API (for automation)
Control plane
Role: exchange routing information, fill routing tables
Technologies: BGP, IS-IS, OSPF, MPLS
Interfaces: CLI (for humans), API (for automation)
Data plane
Role: packet forwarding
Technologies: ASIC, NPU, DPDK, VPP
Interfaces: API (often vendor-specific)

A typical router operates on Data plane and Control plane levels, and provides a management plane interface for network automation systems. Some routers (e.g. route servers) don't have a Data plane and operate only on Control plane level.

Similarly, the SR-TE controller is a Control-plane element. It runs routing protocols (such as BGP), and calculates routing information for traffic engineering.

Unfortunately, many SR-TE controllers currently present on the market have been designed as Network Management Systems. This is a fundamental flaw of those implementations.

1. Management plane protocols have not been designed as critical for network functionality, so convergence times after failures are slow and not guaranteed.
2. Most management systems use GUI - which looks nice if you want to draw a graph of network resource utilization but it is easy to get lost in it when troubleshooting a time-sensitive routing problem.
3. Management plane protocols are poorly standardized and don't interwork between different implementations as well as BGP.
4. Poor standardization also means there are few network engineers familiar with those protocols, which further increases time to fix any issue.

Routing protocols, on the other hand, are far better understood, standardized and have been proven reliable in mission-critical environments. Most network engineers are familiar with the industry-standard CLI format and can quickly isolate any routing problem when required.

Decades of evolution of network devices have brought us to the combination of CLI (for humans) and API (for automation systems).

The next generation controller

The idea behind Traffic Dictator is to make an SR-TE controller that focuses just on being a good controller, instead of trying to combine all network management platforms in one. It might not have fancy interfaces with many icons and bright colors, but it is easy to understand and does the main job: Traffic Engineering.

Industry standard CLI makes it easy for any network engineer to configure Traffic Dictator, and the API lets you integrate it with network automation tools.

BGP SR-TE is chosen as a preferred method of policy advertisement to routers, because it minimizes the amount of required sessions and leverages the existing BGP infrastructure to distribute policies.

BGP Labeled Unicast serves as an alternative method to advertise TE policies to routers that don’t support BGP SR-TE.

Multi-domain capability allows the operator to build SR-TE policies across multiple IGP domains, allowing to scale the network to many thousands of routers and isolate different areas from IGP instability. Thanks to BGP Link-State, Traffic Dictator can have the full view of all IGP topologies and build the optimal path that satisfies all constraints.

It is also possible to interconnect different IGP domains with BGP-only links, such as in Inter-AS MPLS VPN designs. Traffic Dictator supports SR-TE policies across IGP domains with BGP links in between.

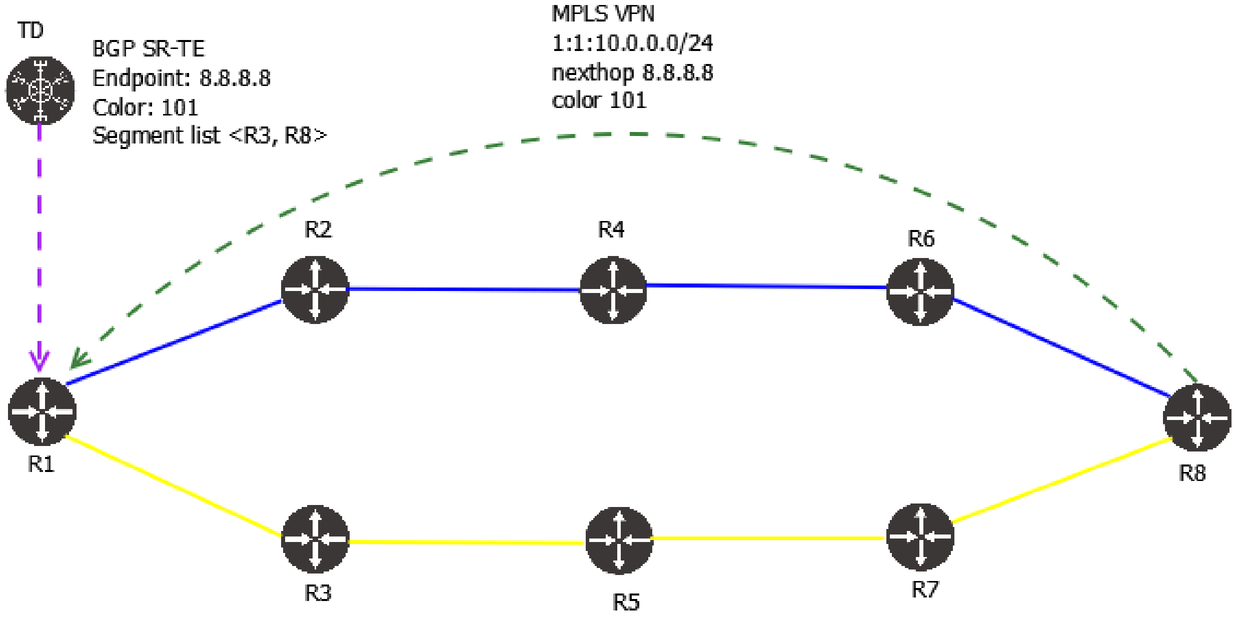
Traffic Dictator revolutionizes Egress Peer Engineering by extending link affinity and bandwidth to egress peers. This way it is possible to use link affinity and bandwidth not only to steer traffic within the network but also to the suitable egress peer.

Null-endpoint policies allow the operator to easily implement bandwidth-aware hot-potato routing, steering traffic to the closest suitable egress peer that matches specified constraints.

Automated Steering with SR-TE

Each SR-TE policy has a color, which can be used to steer traffic into that policy. A BGP route with color extended community matching the SR-TE policy color will be mapped to that policy.

Consider the topology:

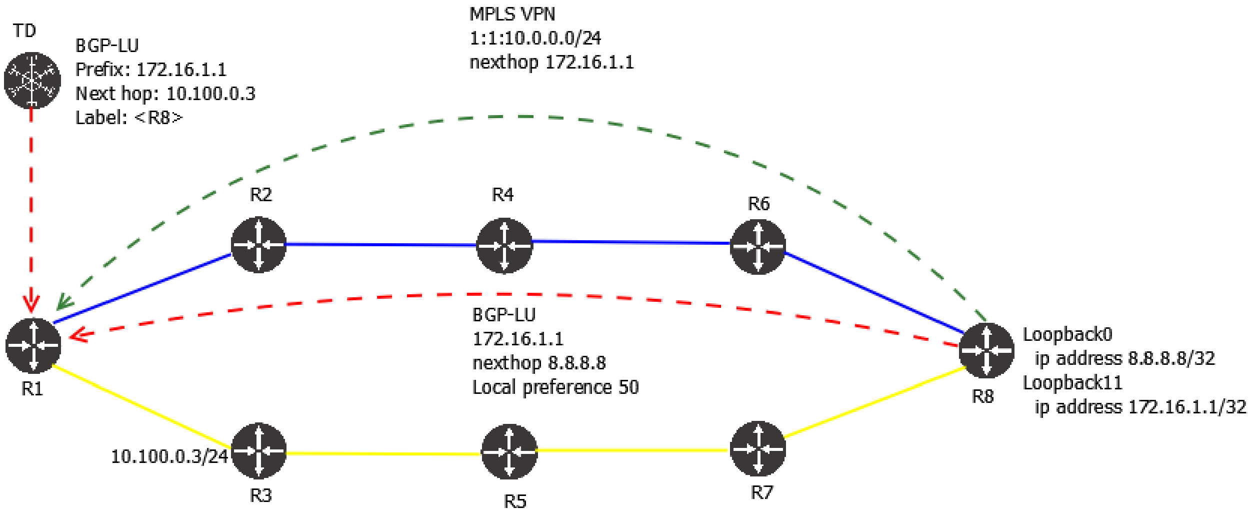


There is an MPLS L3 VPN service between R1 and R8. Design requirements are that traffic for 10.0.0.0/24 is sent strictly over the yellow links. Traffic Dictator calculates an SR-TE policy with color 101, and L3 VPN routes matching endpoint 8.8.8.8 and color 101 will be mapped to that policy.

Emulating Automated Steering with legacy and budget routers

One of the goals of Traffic Dictator is to lower the entry barrier for Traffic Engineering, so that network operators are not locked in with expensive routers from a handful of big vendors, but have a wider choice of various routers available to them, including whitebox routers and open source routing implementations.

Some of these devices don't support BGP SR-TE or PCEP, but almost any router that supports basic MPLS, also supports BGP Labeled Unicast. This allows network operator to use BGP-LU to install Traffic Engineering policies. BGP-LU doesn't have a concept of "color" like SR-TE, but Traffic Dictator offers an option of "service-loopback" to emulate the SR-TE color functionality with BGP-LU.



In this topology, R8 has a "service-loopback" 172.16.1.1. It is not advertised into IGP but advertised into BGP-LU with nexthop of 8.8.8.8 which is the main R8 loopback reachable via IGP. R8 advertises 172.16.1.1 to R1 with a lower local preference, as a backup route to preserve best effort path if the controller fails. When advertising the L3 VPN route for 10.0.0.0/24, R8 sets nexthop to 172.16.1.1.

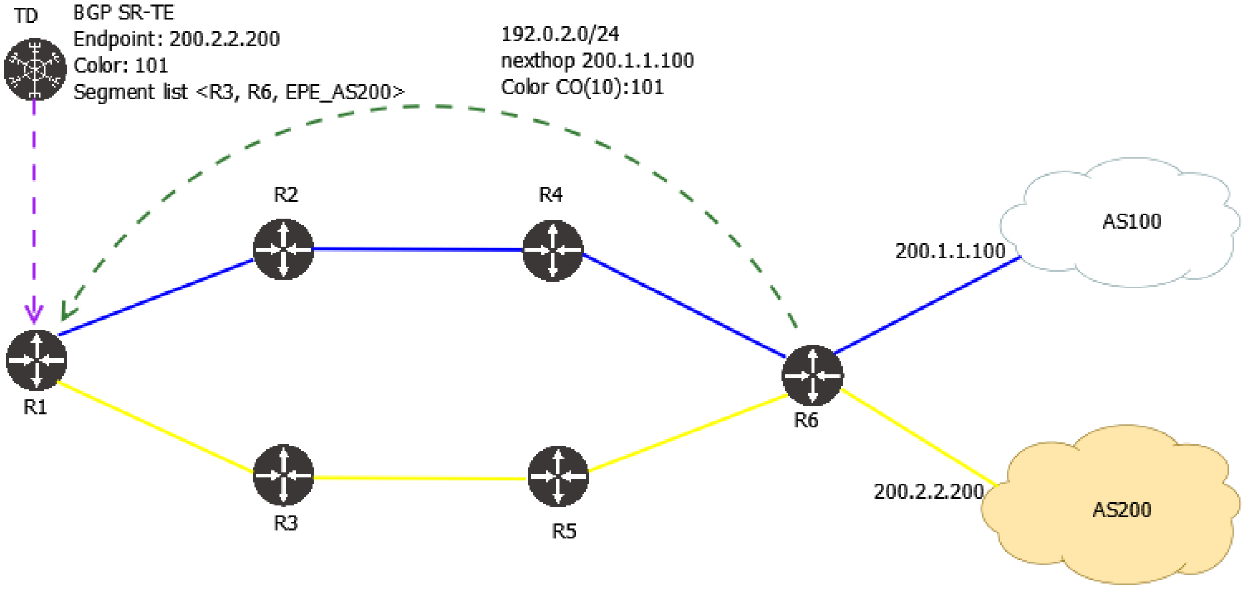
Traffic Dictator sends a BGP-LU route to R1, with prefix 172.16.1.1, and nexthop of R3 interface address. This way, the same behaviour as automated steering can be achieved even if routers do not support BGP SR-TE.

Note that label stack in this case will be just <R8> and not <R3, R8> - this is because unlike SR-TE, where the first label is used by router for nexthop resolution, BGP-LU route already has an IP nexthop so there is no need to send a label of a directly connected router.

Egress Peer Engineering

Traffic Dictator introduces a lot of innovative capabilities around Egress Peer Engineering. It lets the operator use bandwidth and affinity constraints not only within the IGP domain, but also for egress peers. This makes possible new network designs where the network becomes aware of egress bandwidth during reconvergence.

Traffic Dictator expands the concept of link affinities and bandwidth reservations and applies it to EPE. Consider the topology:



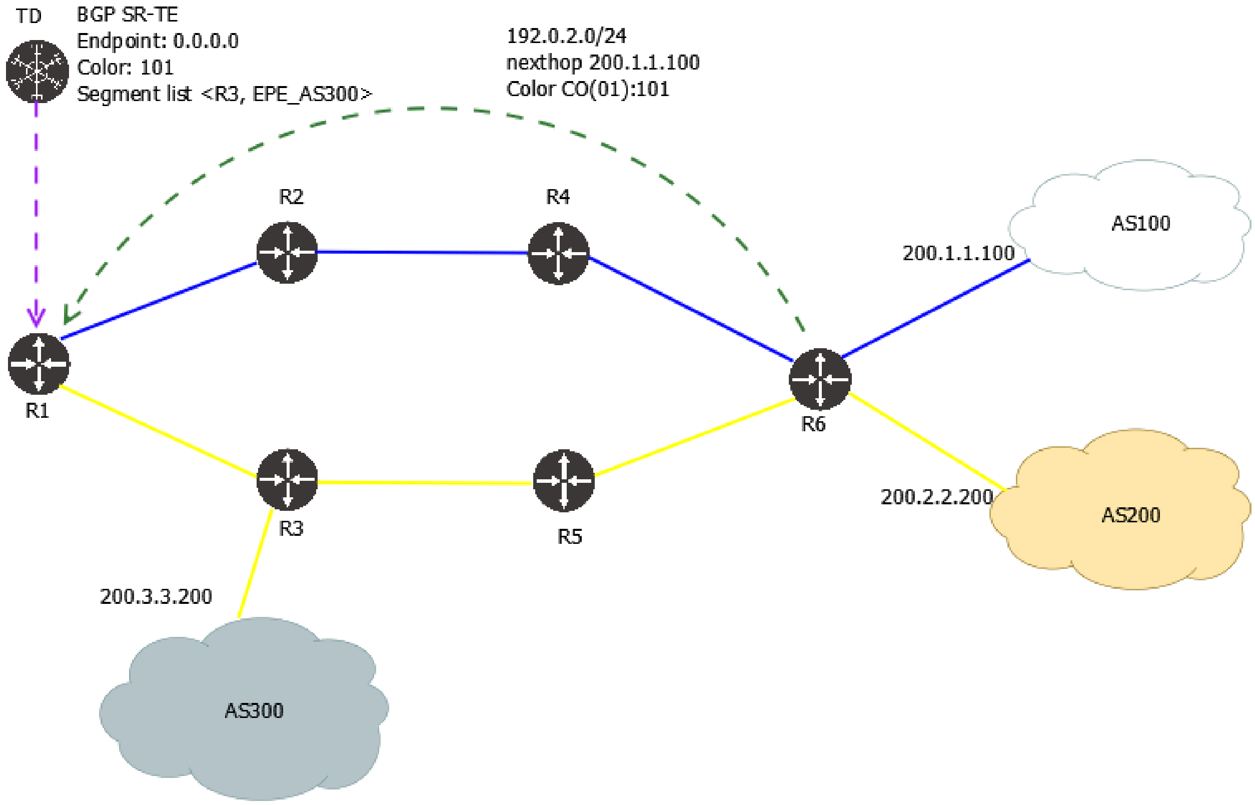
Prefix 192.0.2.0 is advertised from AS100 and AS200. The route via AS100 is preferred due to BGP policies on R6, but there is a requirement to steer this particular prefix via yellow links and AS200.

Traffic Dictator builds an SR-TE policy to AS200 using SID list of R3 and R6 node SID, and R6 EPE label towards AS200.

SR-TE allows the operator to set CO bits in color extended community to ignore BGP nexthop and map prefix to SR-TE policy only based on color (also known as color-only steering). R6 advertises prefix with color CO(10):101 to R1. CO(10) means any endpoint steering. Since there is no SR-TE policy matching nexthop 200.1.1.100 AND color 101, R1 ignores nexthop and maps traffic towards 192.0.2.0/24 to the SR-TE policy via AS200.

EPE with Null Endpoint

Traffic Dictator supports null endpoints with SR-TE. This is useful for hot-potato routing design: when the operator wants to send traffic to the closest suitable exit rather than forwarding through the network to the most optimal exit.



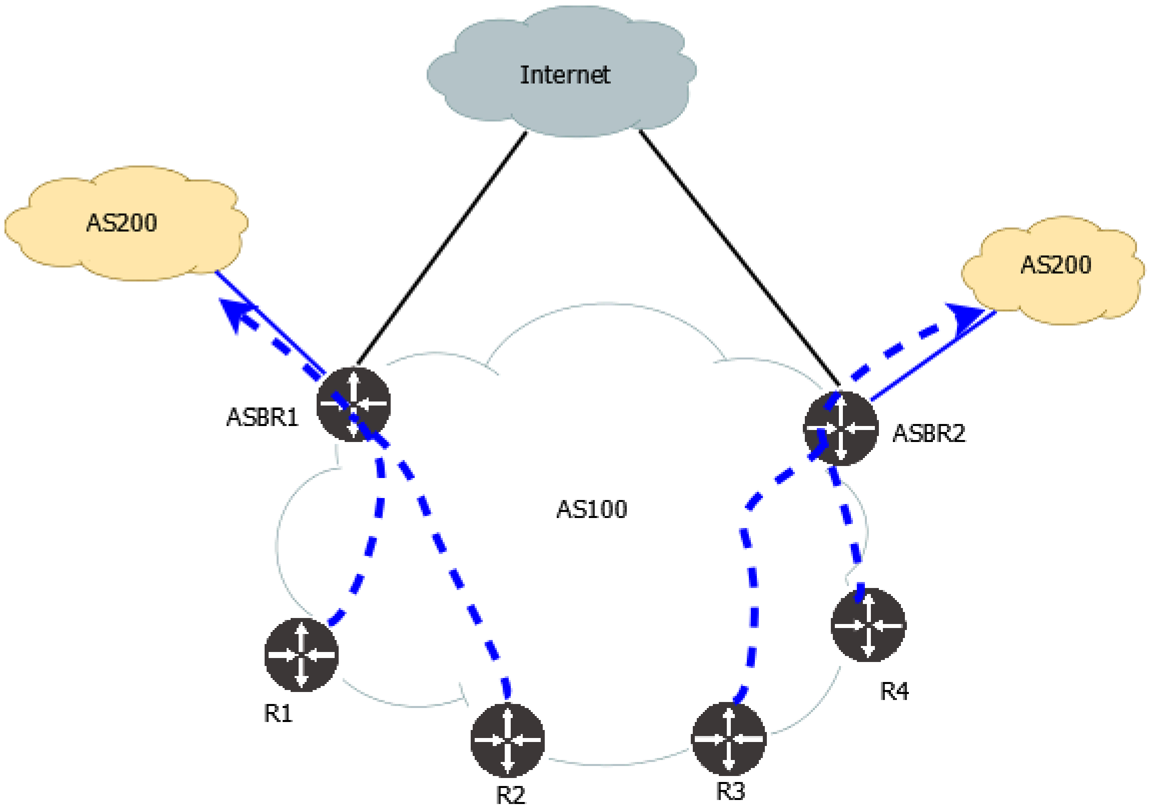
In this topology, there is a requirement to steer traffic to 192.0.2.0/24 using only yellow links, but send it to the closest exit to R1. The operator configures a policy with endpoint 0.0.0.0 color 101 and sets color community on BGP prefix to CO(01):101. This means null-endpoint steering to color 101.

Given endpoint 0.0.0.0 and constraint <yellow only>, Traffic Dictator finds 2 suitable exit points: <R3, AS300> and <R6, AS200>. After evaluating both of them from R1 perspective, <R3, AS300> is preferred due to lower IGP metric. Therefore, TD sends a segment list of R3 node SID and R3's EPE label towards AS300.

While EPE with Null endpoint is a powerful mechanism, it is essential that the network designer creates a good design for color-to-egress peer mapping, such a color per peering partner, or a color for BGP full view, to avoid sending traffic to an AS which doesn't own a given prefix.

Bandwidth-aware EPE

Traffic Dictator also extends bandwidth reservations to EPE. Consider the following topology:

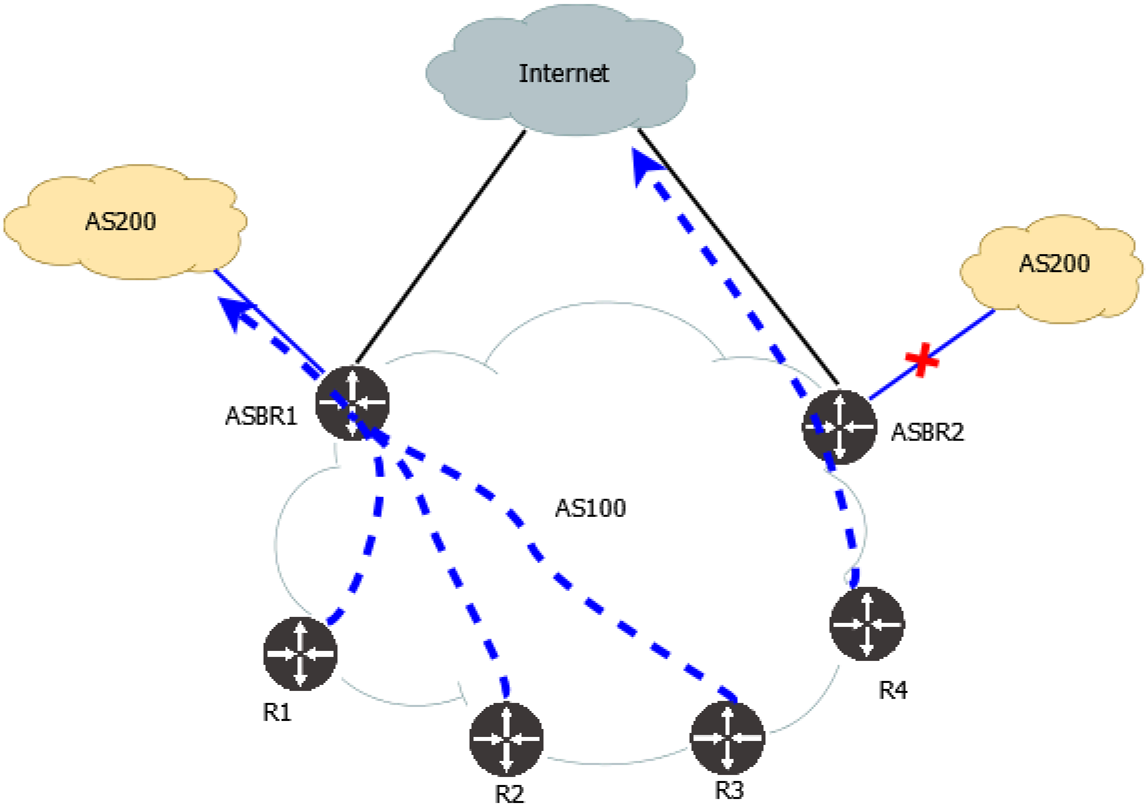


AS100 has 2 peerings with AS200, each peering is over 100 Gbps link. R1, R2, R3, R4 each send 30 Gbps of traffic towards AS200, this traffic is load balanced across the 2 peerings.

AS200 prefixes are also available via the Internet uplink but the operator set higher local preference to private peerings because they are cheaper.

If one of the private peerings fails, all 120 Gbps of traffic will be sent over the remaining private peering, creating congestion and leading to poor network performance and user experience.

Now see what happens with bandwidth-aware EPE:



Traffic Dictator has SR-TE policies for all 4 ingress routers with Null endpoint, and the preferred candidate path via AS200 (blue links). The second candidate path routes traffic via the Internet uplink.

Once the link between ASBR2 and AS200 fails, Traffic Dictator will reroute traffic via the remaining private peering, but since there is not enough bandwidth, the primary candidate path for one of the policies will fail and traffic will be sent over the Internet.

Pure EPE without Segment Routing

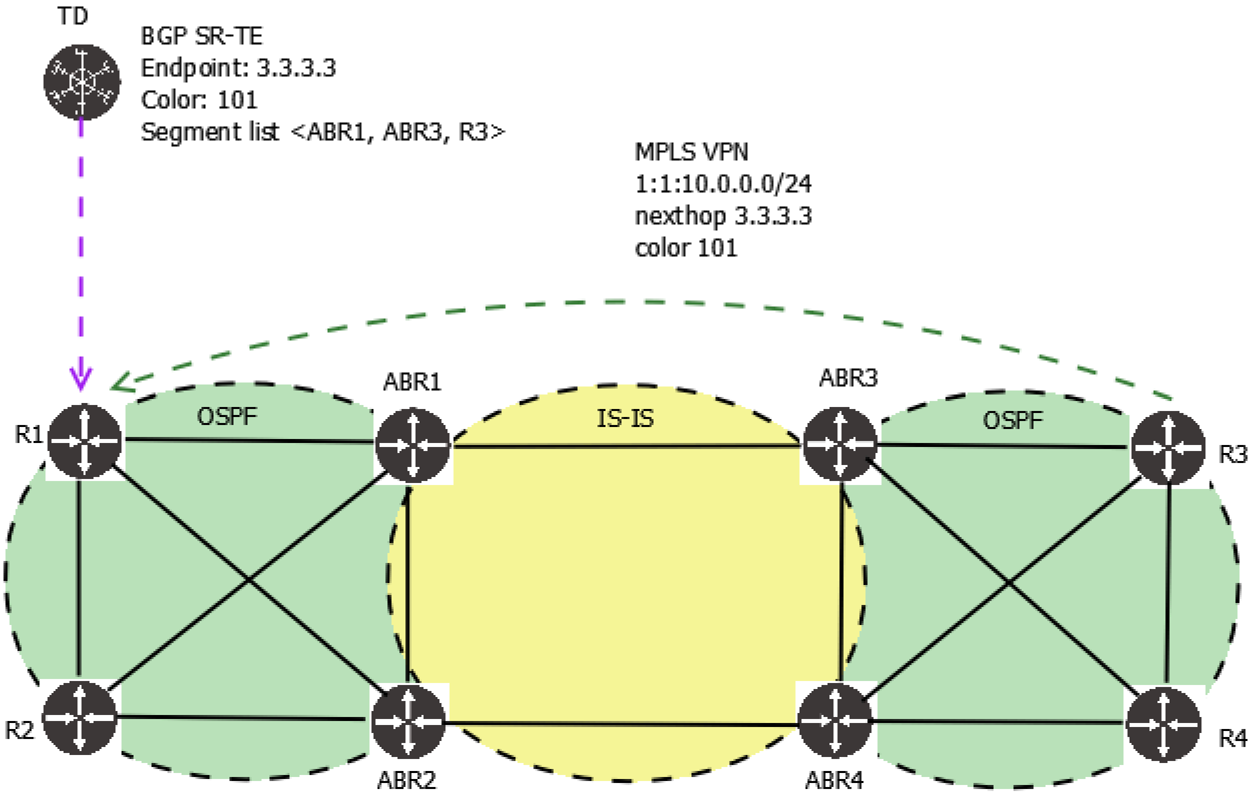
While Traffic Dictator has been primarily designed to work with Segment Routing, it is possible to use Egress Peer Engineering with other MPLS control planes such as LDP or RSVP. In this case, there is no need to advertise IGP topology to Traffic Dictator via BGP-LS; only EPE labels are advertised using BGP-LU or BGP-LS. Then the operator can configure EPE-only policies that will only affect egress peerings but use the path provided by existing MPLS control plane to reach the egress ASBR.

Multi-domain Traffic Engineering

In large networks, sometimes there is a need to split the network into multiple domains, for reasons such as:

- Better scalability
- Isolating network segments from local IGP instability
- Connecting different IGP protocols from various acquired networks

Typically such designs include BGP-LU to connect different IGP domains. Another option is route redistribution between domains, but it is rarely used because it is prone to errors and also defeats the purpose of IGP domain isolation.



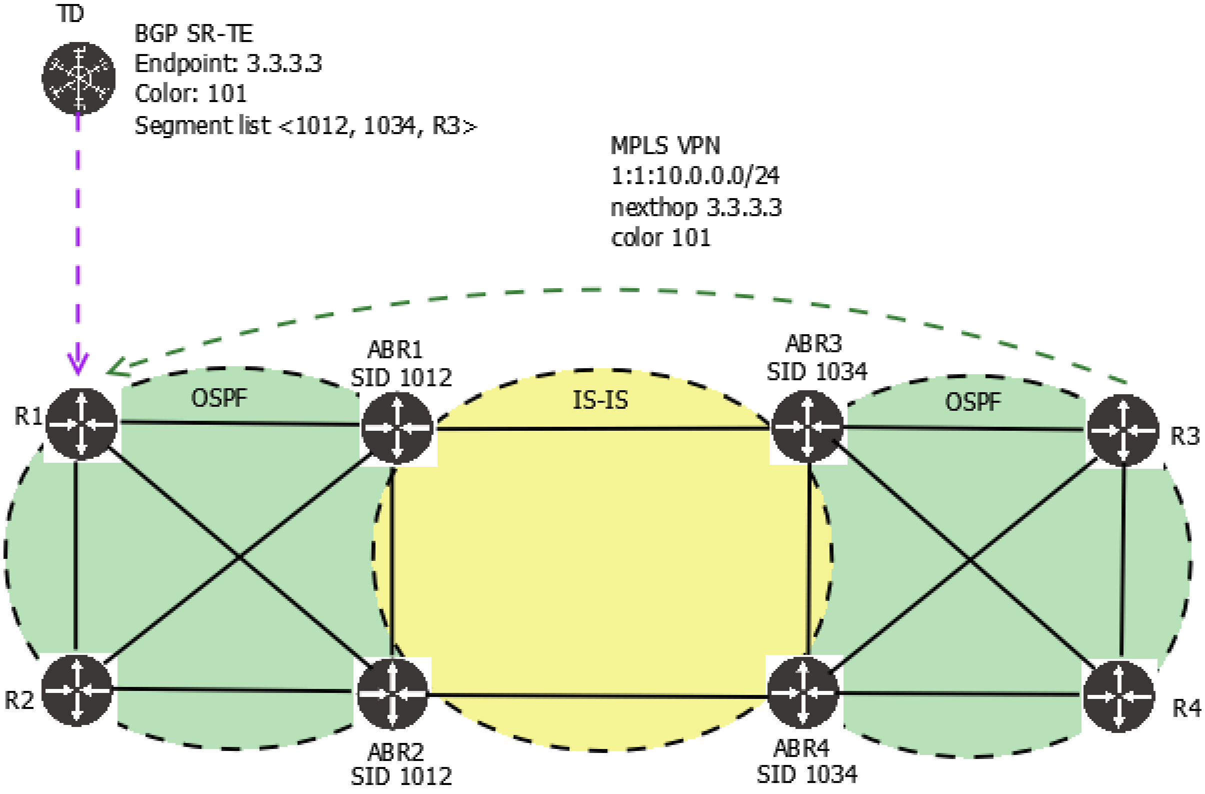
In this example, Traffic Dictator has visibility of all 3 IGP domains. This way it can compute an end-to-end SR-TE policy, including link affinity and bandwidth constraints.

Note the label stack: Traffic Dictator doesn't know whether there is route redistribution or not (in most designs there will be no redistribution), so it will use SID of every border router in the segment list.

Anycast SID

In this design, anycast SID works great as it provides load balancing, redundancy, saves TCAM space on routers and further isolates routing from local IGP instability.

Consider the same topology, bur ABR1 and ABR2 share an anycast SID, and so do ABR3 and ABR4.

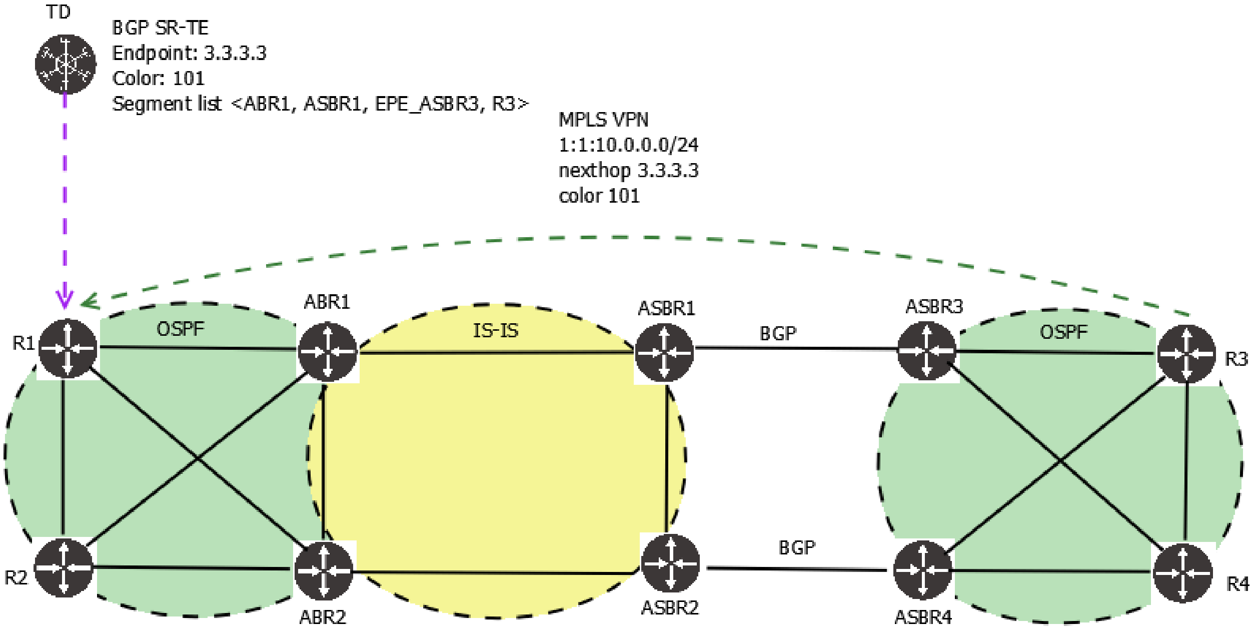


Traffic Dictator will try to use anycast SID in SR-TE policies whenever possible, so traffic will be load balanced across both ABR pairs.

Anycast SID is a preferred design method with Segment Routing and is a significant advantage over RSVP-TE.

BGP-only links and Inter-AS policies

Consider the following topology:

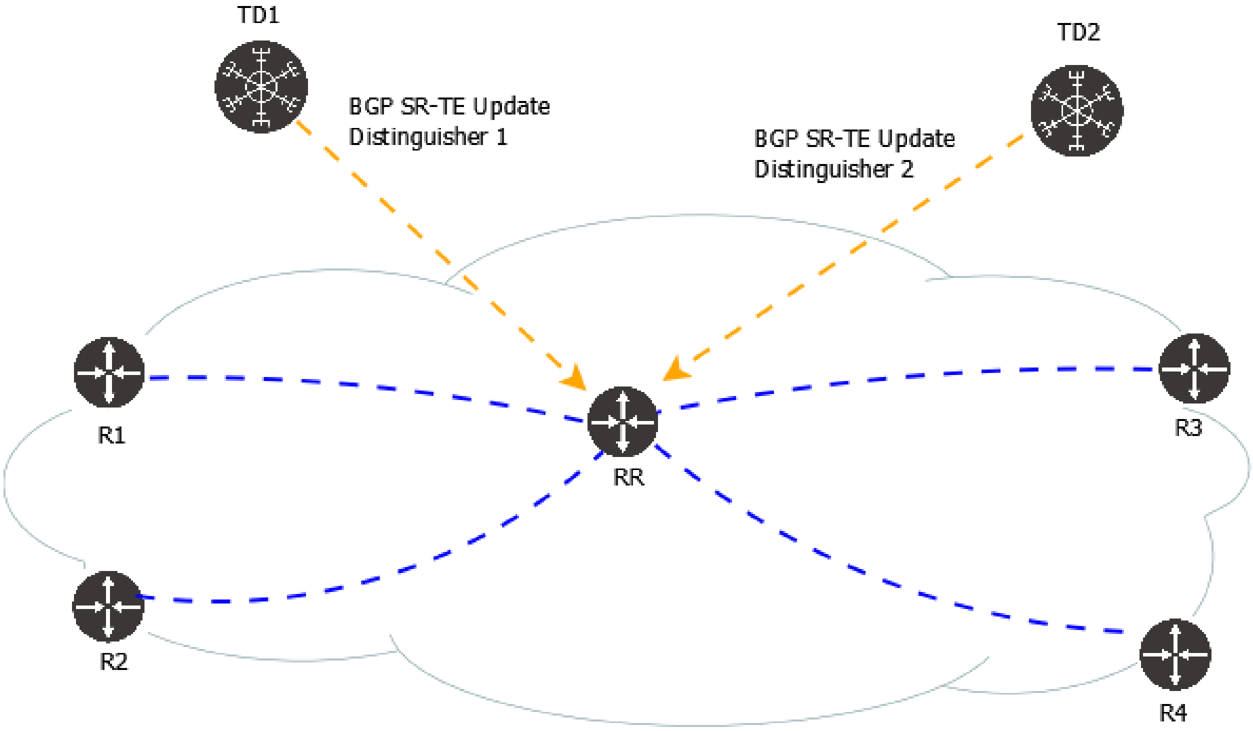


Unlike the previous example, now there are BGP-only links between different IGP domains. Traffic Dictator leverages EPE functionality to build paths in such topologies. It will use the EPE label of the link between ASBR1 and ASBR3 and then calculate the path within the IGP topology.

Such policies can also use affinity and bandwidth constraints.

Redundancy

BGP SR-TE allows an easy and scalable way to design controller redundancy. Multiple controllers advertise SR-TE policies to route-reflectors using different SR-TE distinguishers. This way, even if one controller dies and its NLRI get withdrawn, all routers will have all policies from the second controller, so there will be no interruption to traffic flows. For extra redundancy, it is possible to setup more than 2 controllers the same way.



Deployment and operations

Traffic Dictator is available as a Docker container and a Virtual machine. It can be deployed on a standalone host, in an existing virtualization or container environment, or on a router that can run containers.

Industry-standard CLI is intuitively understood by any network engineer familiar with routers, and makes the product easy to use for PoC testing, studying and troubleshooting.

HTTP API is available for network automation and integration with network management systems.

Further information

A restricted version of Traffic Dictator is available for evaluation, PoC testing, studying and other non-commercial purposes. The limit is that it allows a maximum of 50 policies without a license.

Download Docker container or a VM at from: https://vegvisir.ie/downloads/

For documentation refer to: https://vegvisir.ie/documentation/

For more details about the product and the company, contact: info@vegvisir.ie

For sales information, contact sales@vegvisir.ie

For technical support, contact: support@vegvisir.ie

Author

Dmytro Shypovalov
Founder
Vegvisir Systems Ltd

dmytro@vegvisir.ie

vegvisir.ie