Proyecto ANTEL/FING: "Diseño de topologías iBGP/MPLS óptimas para la Red Internacional de ANTEL"

ANTEL

Oscar Zagarzazú (sponsor)
Pablo Cuello (líder)
José Restaino (técnico)

FING

Claudio Risso (tutor)
Eduardo Grampín (co-tutor)
Cristina Mayr (PhD. Student)

Foro Técnico de LACNIC (FTL2020)

Mayo 4–8, 2020 (Cali, Colombia)

antel

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

FACULTAD DE INGENIERIA

Escalabilidad del Protocolo BGP

Internet es un arreglo de sistemas autónomos (ASes) que comparten sus rutas a través de BGP (Border Gateway Protocol).

La selección de rutas es por cada router, y se basa en el BGP Path Selection Algorithm, que incluye:

- Información administrativa (weight, local preference, MED);
- Operativa (local router, AS-path, prioridad de protocolos);
- E información del IGP, ya que de empatar en los pasos previos, el gateway más cercano ha de ser elegido siguiendo la métrica interna.

En un overlay iBGP full-mesh, cada router recibe todos los updates importantes y desempata con su propia métrica IGP. El mecanismo confiere optimalidad (i.e. full-mesh optimality), lo que a su vez provee consistencia.

El problema es de escalabilidad con el número de routers en el AS, ya que desde cada adyacencia eBGP pueden llegar hasta 900k updates IPv4, y el número crece sostenidamente.

Reflectores de Rutas en BGP

Una alternativa escalable para el overlay de BGP es la *reflexión de rutas*. En ella, uno o más routers son designados como Router Reflector (RRs), mientras que los restantes routers son clientes de los reflectores.

Los clientes conectan a uno o más RRs, que se suponen full-mesh interconectados. El path selection algorithm corre como antes, a excepción de la regla de *split-horizon*, que ahora es violada por los reflectores permitiendo relays de updates entre routers del AS.

El nuevo problema es que lo RRs eligen su update óptimo, de acuerdo a su propia ubicación en la red. Para muchos clientes, esa decisión puede ser diferente de la que habrían tomado de disponer de toda la información.

El objetivo principal de este proyecto fue diseñar un overlay iBGP full-mesh óptimo y resiliente para la Red Internacional de ANTEL, con la menor cantidad de RRs y sesiones. El mismo está a su vez coordinado con el forwarding MPLS.

Optimal Router Reflector Topology Design

El problema de diseñar un overlay iBGP *full-mesh óptimo* es actualmente y ha sido históricamente objeto de estudio. En general el problema admite varias versiones y es computacionalmente duro (NP-Hard).

Se suele usar heurísticas (BGPSep, BGPSep_D, BGPSep_S y Zhang) o fuerza bruta (BGP-ORR) para resolver el problema.

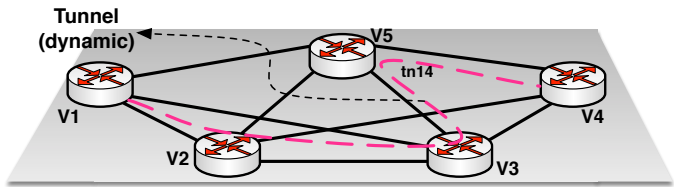
Este equipo ha desarrollado un framework denominado *Optimal Router Reflector Topology Design*. Las formulaciones son exactas (optimización combinatoria). Se han elaborado variantes según:

- Sólo los routers internos son elegibles como RRs
- Adicionalmente, la solución debe preservar la optimalidad ante cualquier falla simple en un nodo o enlace (resiliencia)
- Cualquier router (incluso los ASBRs) puede ser reflector, sosteniendo la resiliencia

Las anteriores son variantes del problema en una red IP pura.

ORRTD en una red IP/MPLS

En la cuarta y última variante implementada, cualquier router es elegible como RR, pero el ruteo IP solamente se da en el borde del AS. Dentro del anterior, el forwarding se basa en MPLS, que es la arquitectura estándar en la actualidad.



En la figura, la decisión IP sería tomada por `V1`, quien determina que `V4` es el *gateway* para un paquete entrante. Las decisiones intermedias no consideran las direcciones IP, y los nodos simplemente usan las etiquetas del frame.

La hibridación de BGP con MPLS previene loops en IP, uno de los potenciales problemas de la suboptimalidad de BGP.

ORRTD en una red IP/MPLS

Otro particular de esta versión es que la resiliencia IP solamente considera pérdidas de adyacencias en los bordes. Los routers se suponen infalibles, por su altísima disponibilidad, ya que en realidad son clusters de nodos carrier-class en datecenters Tier-3/4.

Por tanto, dentro del AS, solamente consideramos fallas en los links, y asumimos que es el failover de MPLS quien las restaura. La resiliencia se diseña aprovisionando un par de caminos físicamente independientes para cada túnel en la red (TE), o asegurando la existencia de un camino no congestionado (LDP).

En esta aplicación, asumimos que la métrica IGP está basada en los delays de propagación, que resultan de los recorridos físicos.

La integración coordinada de los overlays iBGP y la ingeniería de tráfico MPLS es otra contribución académica de este trabajo.

El tomar como referencia la red internacional de ANTEL es el resultado práctico.

Objetivos de la Ingeniería de Tráfico

Label Distribution Protocol (LDP) es el mecanismo más simple para señalizar caminos en MPLS y se basa en replicar los caminos que el ruteo IP puro hubiera elegido siguiendo la métrica del IGP.

Aunque es conocido por su limitada eficiencia en el uso de los recursos, LDP sigue siendo popular por su simplicidad y el paralelismo con el ruteo IP clásico. BGP y los túneles LDP están alineados, ya que ambos usan la misma métrica.

Complementariamente, este trabajo exploró las ventajas de usar ingeniería de tráfico optimizada, al elegir caminos físicamente independientes (primary y secondary paths), administrativamente seteados (señalizados con RSVP-TE) para cumplir con un conjunto de restricciones de Quality of Service (QoS).

En particular, buscando cumplir con límites a los delays entre países, y minimizando la congestión ante cualquier falla física.

El Proceso de Diseño Seguido

El proceso seguido para diseñar los overlays iBGP y MPLS es el detallado a continuación:

- Se toma una captura completa de los updates de la red, y se filtran siguiendo el BGP Path Selection Algorithm.
- Los prefijos resultantes son post-filtrados para descartar aquellos que no habrían sido usados debido a la presencia de otro más específico.
- Los prefijos que hayan sobrevivido se agrupan en clases según la combinación específica de ASBRs que los retransmitirían. Esto reduce sensiblemente la complejidad del problema.
- Se calcula el overlay iBGP óptimo para esas clases.
- El tráfico de cada clase es estimado usando estadísticas por fuente de la red (snmp y netflow).

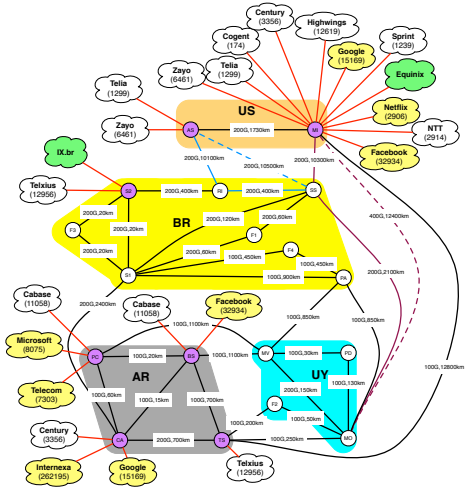
El Proceso de Diseño Seguido

- La estimación del tráfico se ajusta por destino, buscando reproducir el tráfico entre países. Se usa un representante de cada clase en un entorno de emulación para que implemente el overlay iBGP previo. El tráfico sobre ese entorno replica el tráfico medio actual (*nominal case*).
- Se emulan las pérdidas de adyacencias para estimar las matrices de tráfico asociadas en el entorno virtual.
- Se calcula el *worst case scenario* tomando como referencia el máximo tráfico entre cada par de nodos de entre todas las fallas. Puede parecer exagerado, pero frecuentemente (e.g. CDNs), los cambios suceden en forma inesperado, sin coordinación, y una red resiliente debe estar preparada para soportarlo, para coexistir con esa realidad.

El Proceso de Diseño Seguido

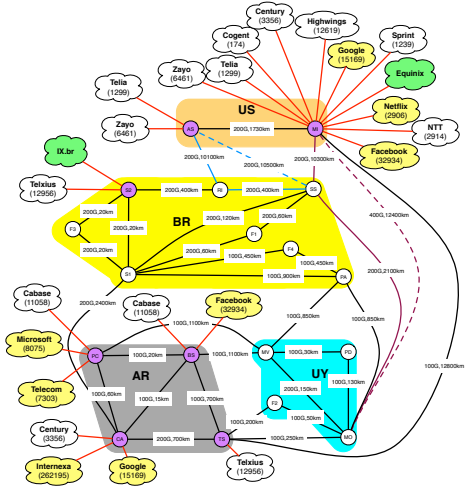
- Los límites de delay entre nodos se definen balanceando los objetivos de diseño con las posibilidades de la red ante las fallas simples potenciales.
- Para mantener consistencia con la que hubiera sido la optimalidad iBGP, es deseable que esos límites estén cerca de los valores IGP. Hay que permitir excepciones cuando eso fuerza la congestión de algunos enlaces.
- Como etapa final, se optimiza la ingeniería de tráfico de la red para encontrar las parejas de caminos independientes de cada túnel, quienes deben cumplir además con los límites de delay, al tiempo que garantizan que no haya congestión incluso luego de cada falla simple en los links.

Ejemplo: un escenario para la Red Internacional de ANTEL



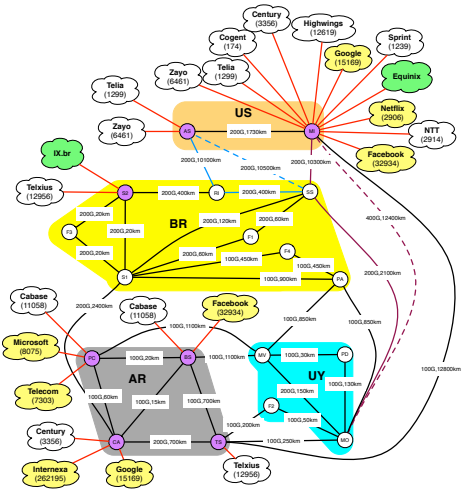
Esta red ficticia recibiría más de 9 millones de updates eBGP de casi 1260 peers IPv4 en cuatro países.

Ejemplo: un escenario para la Red Internacional de ANTEL



La selección de BGP los reduce a unos 800 mil prefijos, mientras que el post-procesamiento los lleva a 743041.

Ejemplo: un escenario para la Red Internacional de ANTEL

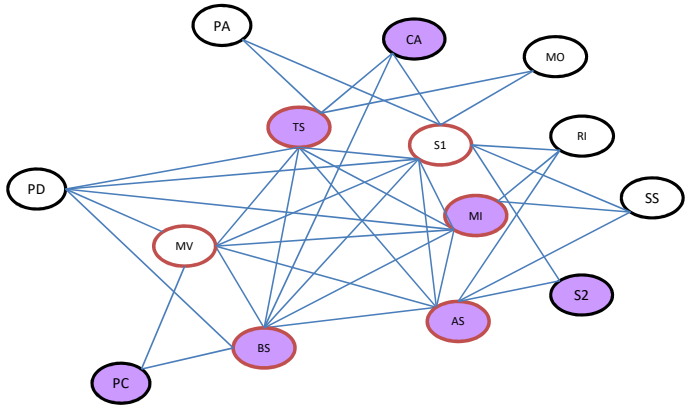


Class Id Identifier	ASBRs codes	Prefixes quantity	Cumulative Percentage
1	TS CA S2 MI	239183	32.2%
2	TS CA S2 AS MI	205359	59.8%
3	MI	53072	67.0%
4	CA	38908	72.2%
5	AS MI	36029	77.1%
6	CA MI	32652	81.4%
7	TS S2 MI	29510	85.4%
8	CA AS MI	26426	89.0%
9	TS S2 AS MI	21228	91.8%
10	TS CA S2	20504	94.6%
11	TS S2	18347	97.1%
12	TS	7914	98.1%
13	PC	4841	98.8%
14	CA S2 MI	3369	99.2%
15	CA S2 AS MI	3083	99.6%
16	S2	462	99.7%
17	AS	427	99.8%
18	CA S2	411	99.8%
19	S2 AS MI	387	99.9%
20	S2 MI	316	99.9%
21	PC CA	258	100.0%
22	TS CA	248	100.0%
23	TS CA S2 AS	48	100.0%
24	BS MI	34	100.0%
25	CA AS	11	100.0%
26	TS PC S2	8	100.0%
27	TS PC CA S2	6	100.0%

Es un hecho a destacar que todos ellos pueden agruparse en 27 clases, y que más del 90% se concentra en las primeras 9 de ellas.

El overlay iBGP óptimo (escenario de ANTEL)

El siguiente es el overlay iBGP óptimo para esa red y esas clases. Los nodos ficticios `F1` a `F4` se han obviado en el esquema.



Hay 6 RRs (en rojo), 22 sesiones cliente-reflector, además de las 15 sesiones entre reflectores, totalizando 37 sesiones, que corresponden a un tercio de las 91 en un full-mesh de 14 nodos.

El overlay iBGP óptimo vs Heurísticas

Comparar contra las heurísticas clásicas no es directo porque ellas no tienen en cuenta el detalle de los prefijos, ni la resiliencia en algunos casos.

Como referencia, usamos una única clase default publicada por todos los bordes sobre la misma topología de ANTEL.

Los resultados son:

	ORRTD	BGPSep	BGPSep_D	BGPSep_S	Zhang
#RRs	3	6	6	7	4
#sessions	23	76	67	68	35

Observar que ORRTD no solamente las supera en la simplicidad del overlay, sino también en que preserva la optimalidad luego de cualquier pérdida de adyacencias en cualquier nodo.

Ingeniería de Tráfico

La matriz de demandas nominal totaliza 346Gbps de tráfico. Con los límites de delay objetivos propuestos, la topología de la figura tiene solución factible.

El número trepa a 495Gbps en el *worst-case scenario* (43% más alto). En este caso no hay solución factible, y la congestión de algunos enlaces puede llegar al 25% ante combinaciones de caídas de links y pérdidas de adyacencias.

El modelo optimización combinatoria para la ingeniería de tráfico, apunta a optimizar el uso de recursos de la red, y cuando no hay solución factible, indica cuáles enlaces son *cuello de botella*.

Un resultado destacable de la instancia es que expandiendo selectivamente sus links, un 10% de costo-en-capacidad adicional consigue una red donde nunca se supera el 60% de las capacidad de los links para el escenario worst-case, esto es, ante las combinaciones en fallas de adyacencias y enlaces.

Resultados (Ingeniería de Tráfico óptima)

node Id1	node Id2	length (km)	capacity (Gbps)	slack of bandwidth
AS	MI	1730	200	50
AS	RI	10100	200	21
AS	SS	10500	200	194
MI	SS	10300	200	126
MI	TS	12800	100	25
MI	MO	12400	400	42
PA	S1	900	100	11
PA	MO	850	100	33
PA	MV	850	100	17
PA	F4	450	100	33
RI	SS	400	200	170
RI	S2	400	200	111
SS	S1	120	200	179
SS	F1	60	200	170
SS	MO	2100	200	194
S1	S2	20	200	65
S1	F1	60	200	170
S1	CA	2400	200	158
S1	F3	20	200	133
S1	F4	450	100	33
S2	F3	20	200	133
BS	CA	15	100	-23
BS	PC	20	100	44
BS	TS	700	100	15
BS	MV	1100	100	12
CA	PC	60	100	50
CA	TS	700	200	112
PC	MV	1100	100	94
TS	MO	250	100	-25
TS	F2	200	100	-25
MO	MV	150	200	-9
MO	PD	130	100	-15
MO	F2	50	100	-25
MV	PD	30	100	-15

node Id1	node Id2	length (km)	capacity (Gbps)	slack of bandwidth absolute	slack of bandwidth relative
AS	MI	1730	200	110	55%
AS	RI	10100	200	81	41%
AS	SS	10500	200	134	67%
MI	SS	10300	**300**	151	50%
MI	TS	12800	100	55	55%
MI	MO	12400	400	280	70%
PA	S1	900	**200**	80	40%
PA	MO	850	**200**	133	67%
PA	MV	850	100	40	40%
PA	F4	450	100	48	48%
RI	SS	400	200	89	45%
RI	S2	400	200	96	48%
SS	S1	120	200	83	42%
SS	F1	60	200	154	77%
SS	MO	2100	200	119	60%
S1	S2	20	200	94	47%
S1	F1	60	200	124	62%
S1	CA	2400	200	96	48%
S1	F3	20	200	110	55%
S1	F4	450	100	48	48%
S2	F3	20	200	110	55%
BS	CA	15	**300**	177	59%
BS	PC	20	100	65	65%
BS	TS	700	**200**	80	40%
BS	MV	1100	**200**	91	46%
CA	PC	60	100	42	42%
CA	TS	700	200	100	50%
PC	MV	1100	100	62	62%
TS	MO	250	**300**	135	45%
TS	F2	200	**300**	142	47%
MO	MV	150	**400**	216	54%
MO	PD	130	**300**	162	54%
MO	F2	50	**300**	142	47%
MV	PD	30	**300**	185	62%

Conclusiones y Trabajo Futuro

Este trabajo muestra la conveniencia de coordinar óptimamente los overlays de BGP-IP y MPLS, y cómo el uso de clases simplifica notablemente el problema.

El tamaño de las instancias permitió el uso de solvers comerciales (IBM ILOG CPLEX(R) Interactive Optimizer version 12.6.3). Instancias más grandes y complejas, especialmente en el diseño de la ingeniería de tráfico, requieren desarrollar heurísticas.

Los resultados muestran que la red del esquema es nominalmente factible, y que un 10% de inversión adicional le permite extender la factibilidad al worst-case, con holguras de 40% en los enlaces.

Finalmente, se simuló la señalización en LDP y se verificó que la red no es factible en ninguno de los escenarios, existiendo múltiples enlaces arriba del 100% de congestión para el escenario de demanda worst-case.

Se necesitan inversiones entre 20-25% adicionales para sostener los objetivos con LDP, lo que (en esta red al menos) constituye una referencia para ese sobrecosto.

Esto verifica la ventaja teórica de la ingeniería de tráfico off-line y centralizada, ante la performance de los protocolos dinámicos.

Publicaciones Asociadas I

C. Mayr, C. Risso, and E. Grampín. "A Combinatorial Optimization Framework for the Design of resilient iBGP Overlays". In: *15th International Conference on the Design of Reliable Communication Networks (DRCN19)*. Coimbra, Portugal: IEEE Xplore, 2019, pp. 6–10. ISBN: 978-1-5386-8461-0. DOI: 10.1109/DRCN.2019.8713744.

C. Mayr, C. Risso, and E. Grampín. "Designing an Optimal and Resilient iBGP Overlay with extended ORRTD". In: *The Fifth International Conference on Machine Learning, Optimization, and Data Science (LOD19)*. Lecture Notes in Computer Science. Siena–Tuscany, Italy: Springer International Publishing, 2019, pp. 409–421. ISBN: 978-3-030-37598-0. DOI: https://doi.org/10.1007/978-3-030-37599-7_34.

Publicaciones Asociadas II


C. Mayr, C. Risso, and E. Grampín. "Optimal Route Reflection Topology Design". In: *Proceedings of the 10th Latin America Networking Conference*. LANC '18. São Paulo, Brazil: Association for Computing Machinery, 2018, pp. 65–72. ISBN: 9781450359221. DOI: 10.1145/3277103.3277124. URL: https://doi.org/10.1145/3277103.3277124.


C. Risso, C. Mayr, and E. Grampín. "A combined BGP and IP/MPLS resilient transit backbone design". In: *11th International Workshop on Resilient Networks Design and Modeling (RNDM19)*. Nicosia, Cyprus, 2019, pp. 1–8. ISBN: 978-1-7281-4698-0. DOI: 10.1109/RNDM48015.2019.8949099.

Por cualquier información adicional, remitirse a: crisso@fing.edu.uy.