Programming networks

Not your standard API





Laurent Vanbever

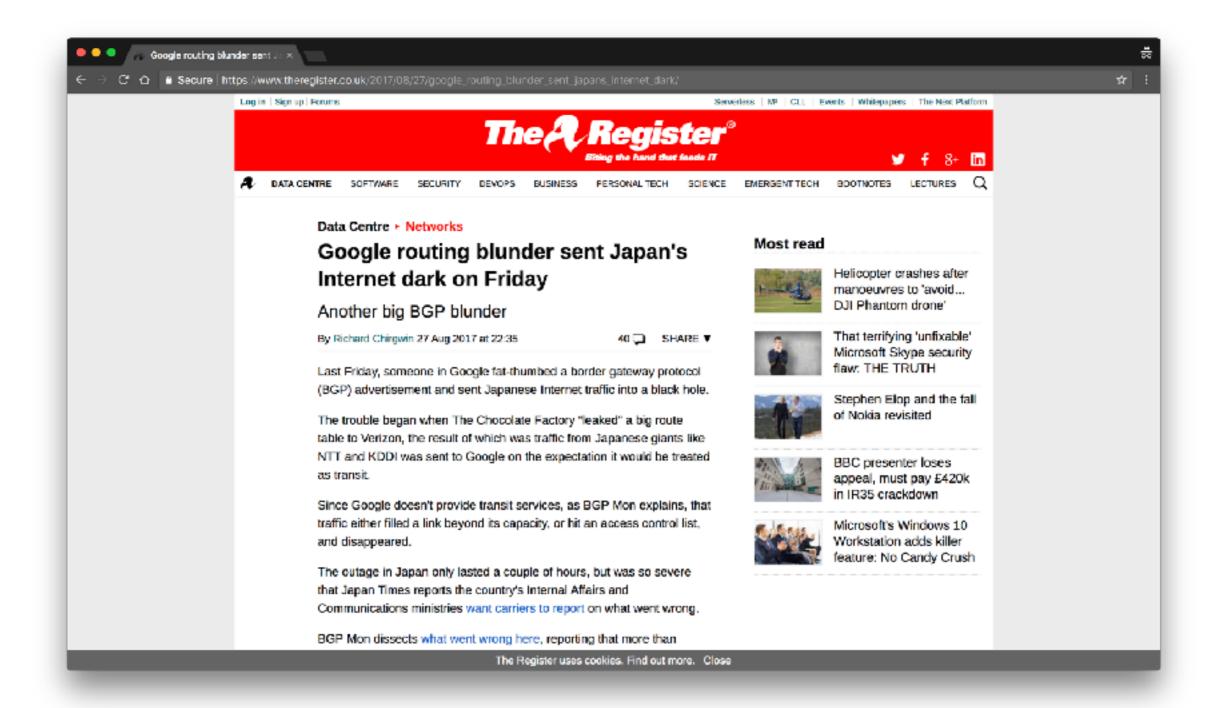
nsg.ee.ethz.ch

NII Shonan Meeting

Mon Feb 26 2018

a couple of hours

a couple of hours??



In August 2017

Someone in Google fat-thumbed a Border Gateway Protocol (BGP) advertisement and sent Japanese Internet traffic into a black hole.

In August 2017

Someone in Google fat-thumbed a
Border Gateway Protocol (BGP) advertisement
and sent Japanese Internet traffic into a black hole.

[...] Traffic from Japanese giants like NTT and KDDI was sent to Google on the expectation it would be treated as transit.

In August 2017

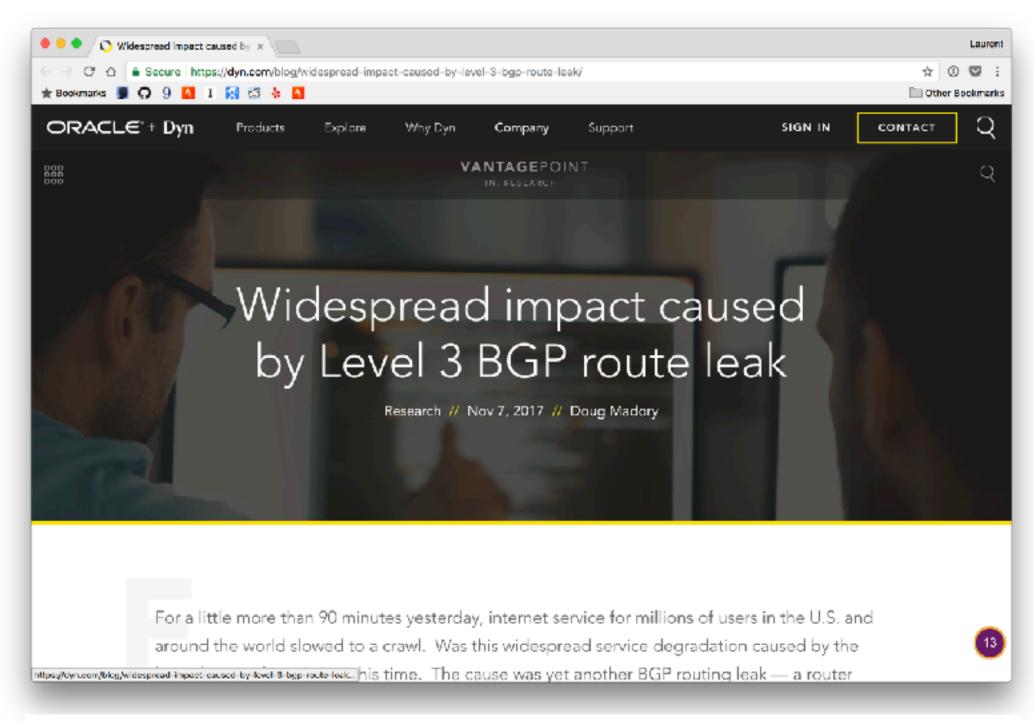
Someone in Google fat-thumbed a Border Gateway Protocol (BGP) advertisement and sent Japanese Internet traffic into a black hole.

[...] Traffic from Japanese giants like NTT and KDDI was sent to Google on the expectation it would be treated as transit.

The outage in Japan only lasted a couple of hours but was so severe that [...] the country's Internal Affairs and Communications ministries want carriers to report on what went wrong.

Another example,

this time from November 2017



https://dyn.com/blog/widespread-impact-caused-by-level-3-bgp-route-leak/

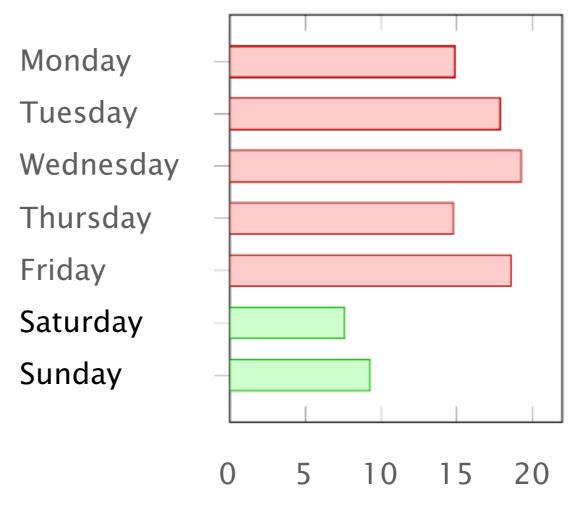
For a little more than 90 minutes [...],

Internet service for millions of users in the U.S. and around the world slowed to a crawl.

The cause was yet another BGP routing leak, a router misconfiguration directing Internet traffic from its intended path to somewhere else.

"Human factors are responsible for 50% to 80% of network outages"

Ironically, this means that the Internet works better during the week-ends...



% of route leaks

source: Job Snijders (NTT)

It's not only about people breaking the Internet...

People destroy their own infrastructure too!



Traders work on the floor of the New York Stock Exchange (NYSE) in July 2015. (Photo by Spencer Platt/Getty Images)

DOWNTIME

UPDATED: "Configuration Issue" Halts Trading on NYSE

The article has been updated with the time trading resumed.

A second update identified the cause of the outage as a "configuration issue."

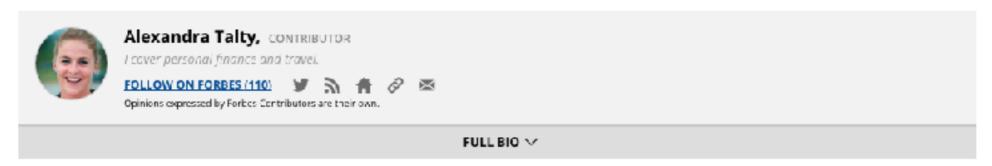
A third update added information about a software update that created the configuration issue.

NYSE network operators identified the culprit of the 3.5 hour outage, blaming the incident on a

"network configuration issue"

JUL 8, 2015 @ 03:36 PM 11,261 VIEWS

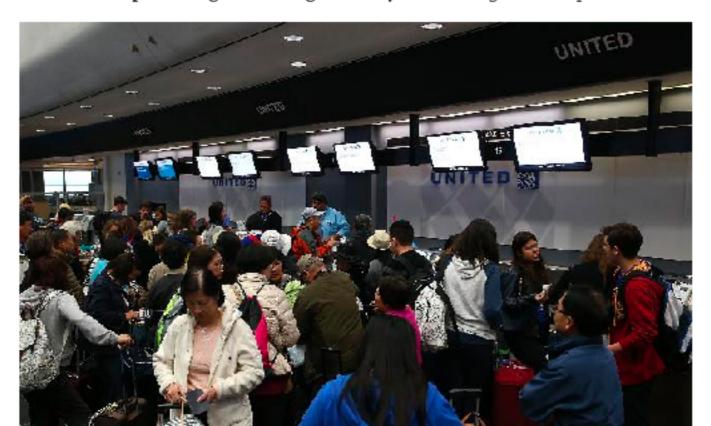
United Airlines Blames Router for Grounded Flights



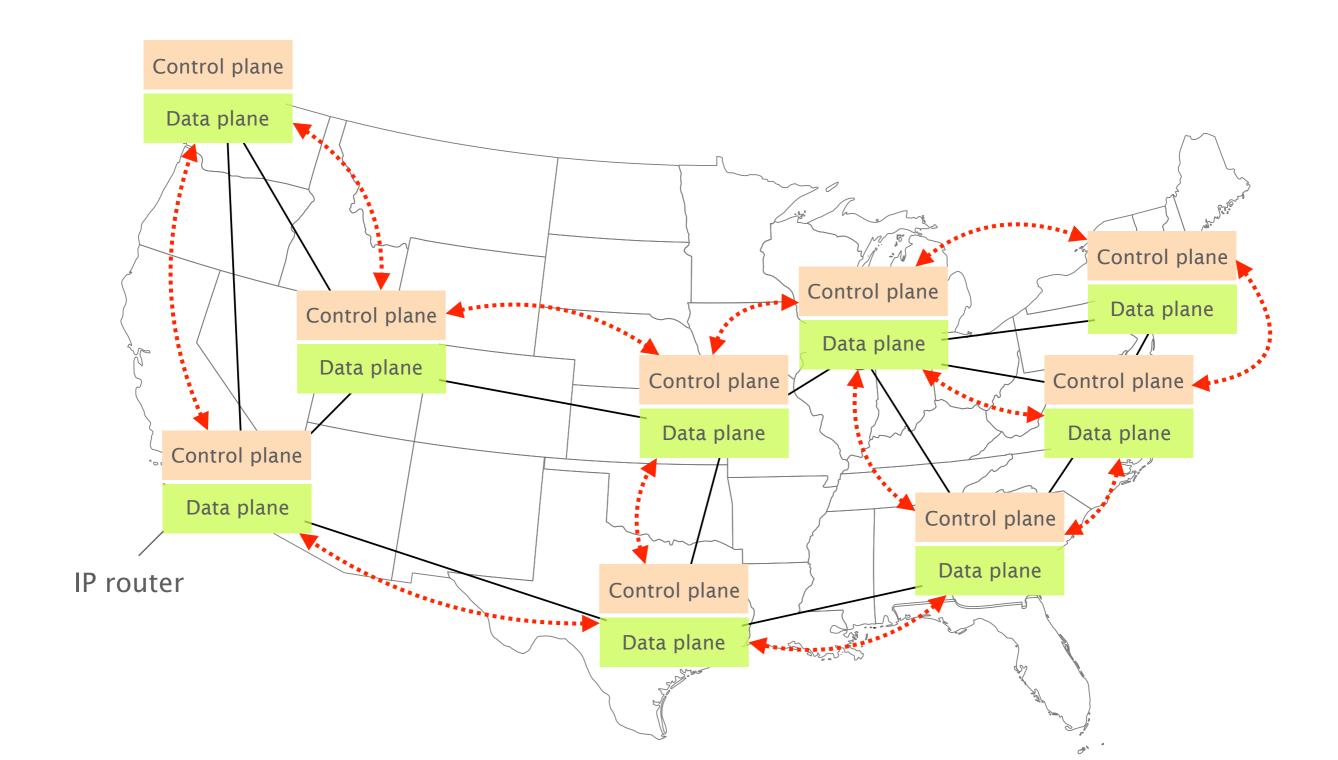
After a computer problem caused nearly two hours of grounded flights for United Airlines this morning and ongoing delays throughout the day, the airline announced the culprit: a faulty router.

Spokeswoman Jennifer Dohm said that the router problem caused "degraded network connectivity," which affected various applications.

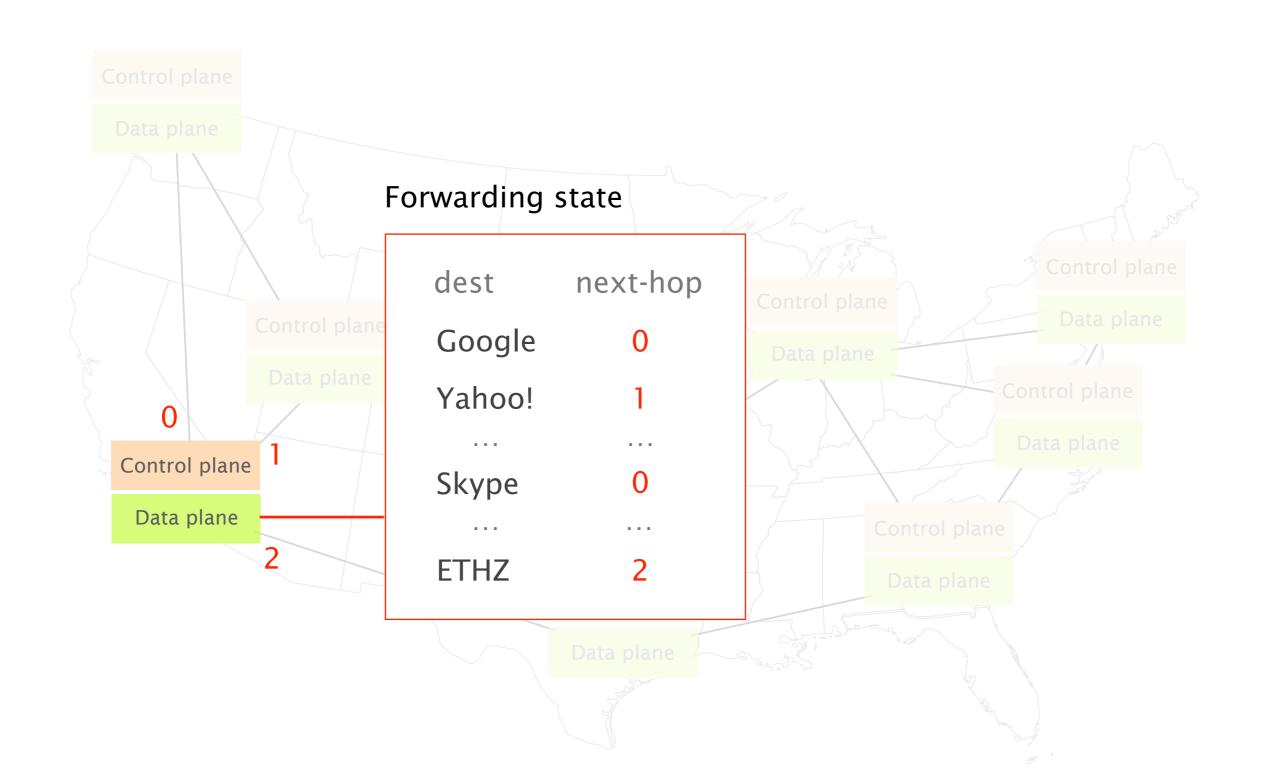
A computer glitch in the airline's reservations system caused the Federal Aviation Administration to impose a groundstop at 8:26 a.m. E.T. Planes that were in the air continued to operate, but all planes on the ground were held. There were reports of agents writing tickets by hand. The ground stop was lifted around 9:47 a.m. ET.



Think of the network as a distributed system running multiple distributed algorithms



These algorithms produce the forwarding state which drives IP traffic to its destination



Operators adapt their network forwarding behavior by configuring each network device individually

Configuring each element is often done manually, using arcane low-level, vendor-specific "languages"

Cisco IOS

```
ip multicast-routing
interface Loopback0
ip address 120.1.7.7 255.255.255.255
ip ospf 1 area 0
interface Ethernet0/0
no ip address
interface Ethernet0/0.17
encapsulation dot1Q 17
ip address 125.1.17.7 255.255.255.0
ip pim bsr-border
ip pim sparse-mode
router ospf 1
router-id 120.1.7.7
redistribute bgp 700 subnets
router bgp 700
neighbor 125.1.17.1 remote-as 100
address-family ipv4
 redistribute ospf 1 match internal external 1 external 2
 neighbor 125.1.17.1 activate
address-family ipv4 multicast
 network 125.1.79.0 mask 255.255.255.0
  redistribute ospf 1 match internal external 1 external 2
```

Juniper JunOS

```
interfaces {
   so-0/0/0 {
        unit 0 {
            family inet {
                address 10.12.1.2/24;
            family mpls;
   ge-0/1/0 {
        vlan-tagging;
        unit 0 {
            vlan-id 100;
            family inet {
                address 10.108.1.1/24;
            family mpls;
        unit 1 {
            vlan-id 200;
            family inet {
                address 10.208.1.1/24;
protocols {
    mpls {
        interface all;
    bgn {
```

A single mistyped line is enough to bring down the entire network

Cisco IOS

```
redistribute bgp 700 subnets — Anything else than 700 creates blackholes family inet {
```

Juniper JunOS

My research goal? Automate!

Remove the need to rely on humans

Develop a complete & sound network controller which can automatically enforce high-level requirements

Develop a complete & sound network controller which can automatically enforce high-level requirements

Twist Control networks running distributed protocols

where the computation of the forwarding state is distributed

Develop a complete & sound network controller which can automatically enforce high-level requirements

Twist

Control networks running distributed protocols

where the computation of the forwarding state is distributed

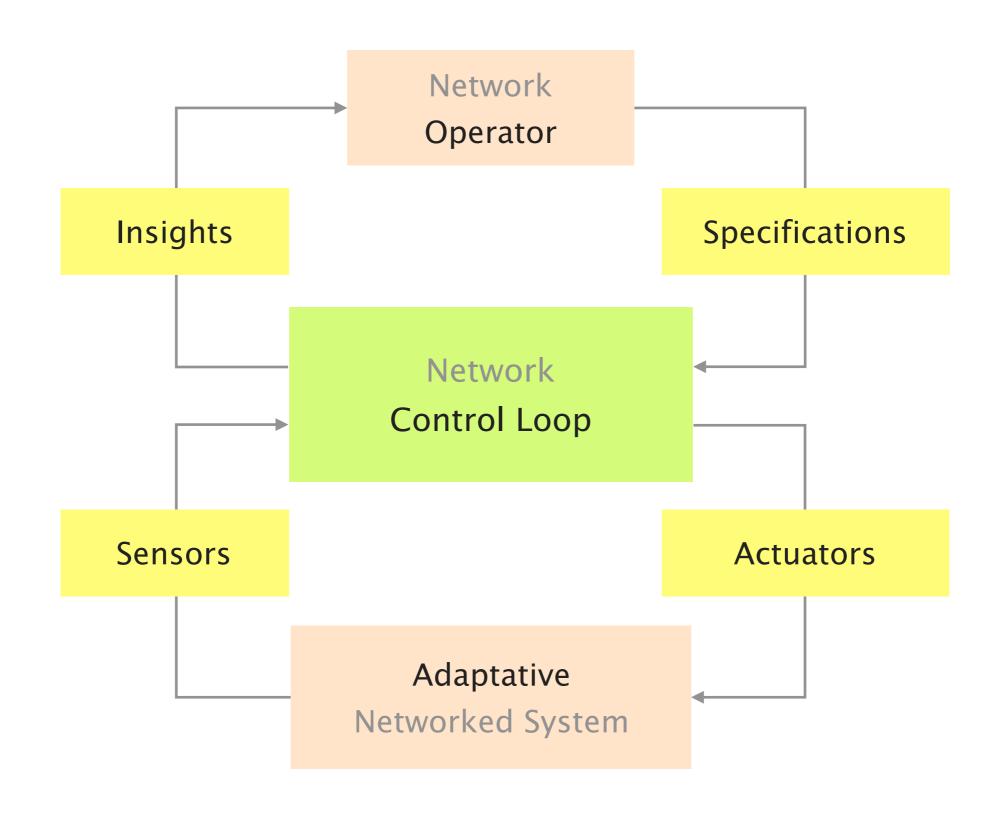
Why?

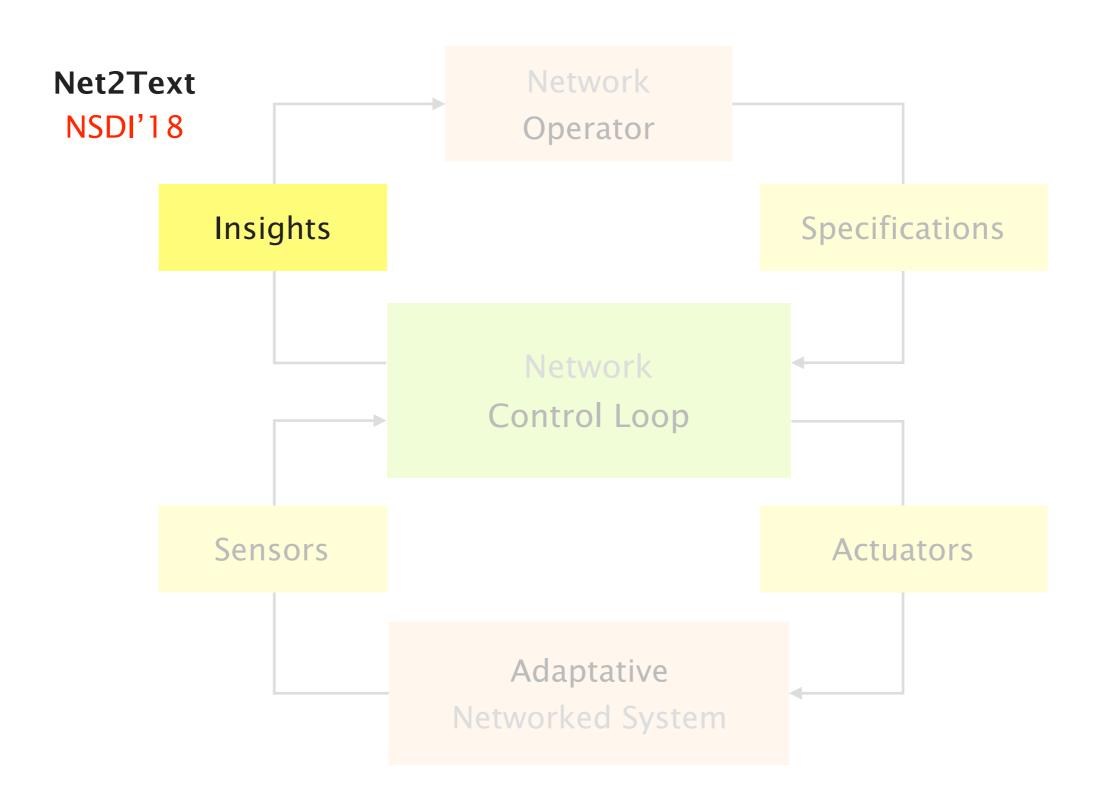
Designing central, scalable and robust control is hard

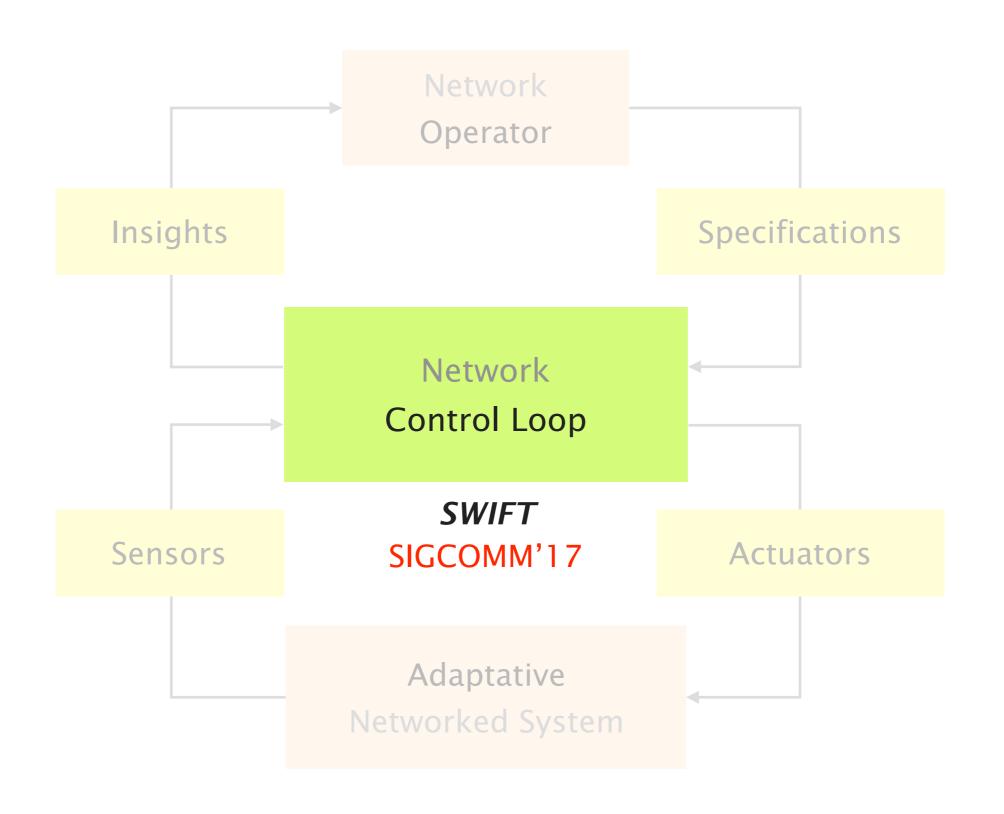
must ensure always-on connectivity to the controller

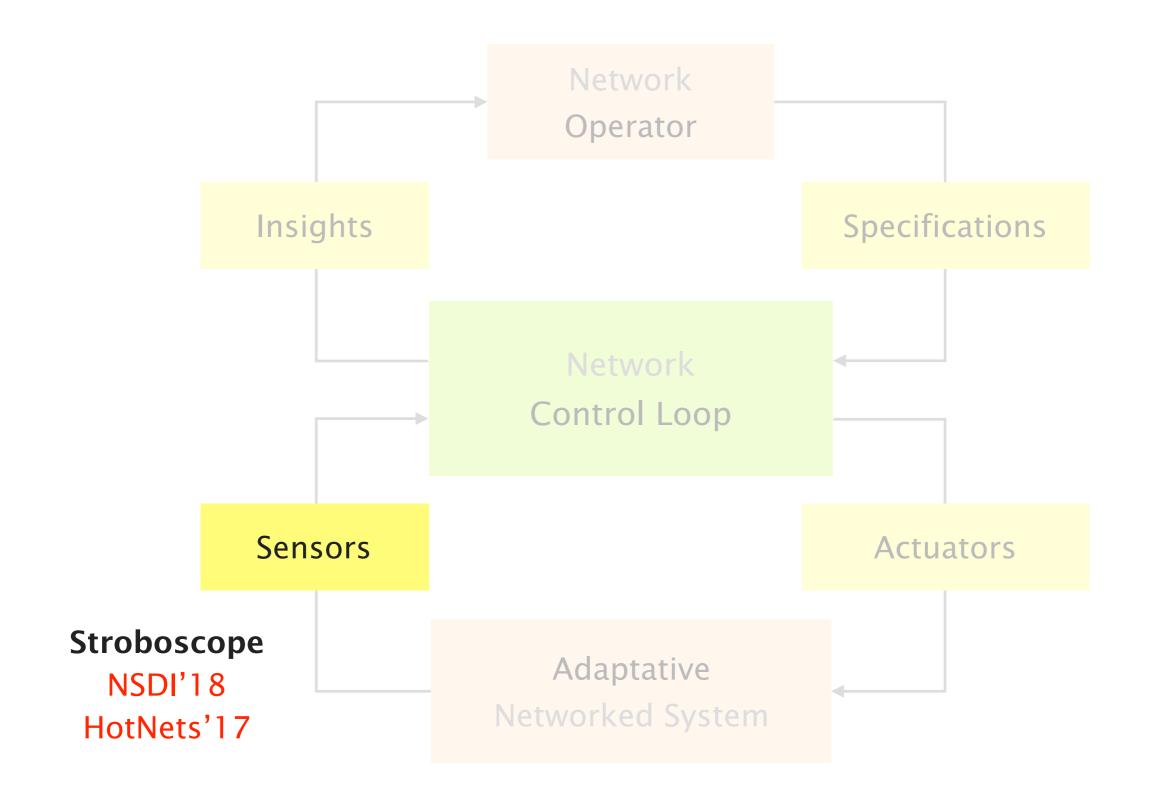
Distributed protocols are still ruling over networks

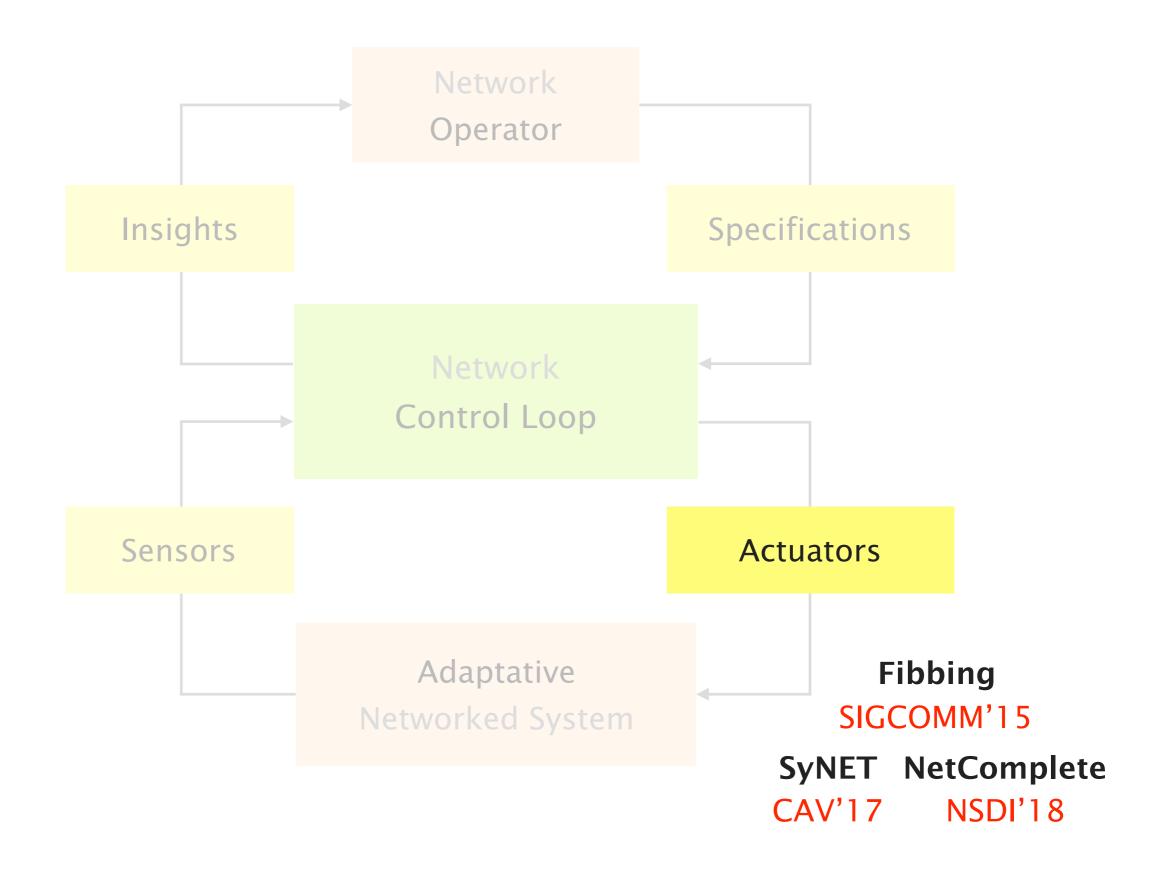
the vast majority of the networks rely on OSPF, BGP, MPLS, ...











How can we control the network-wide forwarding state produced by distributed protocols?

How can we control the network-wide forwarding state produced by distributed protocols?

What are our knobs?

The network-wide forwarding state depends on three parameters

Network-wide Forwarding state

Topology + Network-wide (fixed) + Network - Network - Network - Environment - Forwarding state

Topology + Network-wide Configuration + Network Environment - Network Forwarding state

links & nodes status routing announcements

Out of these three parameters, two can be controlled



links & nodes status

routing announcements

Given a forwarding state we want to program, we therefore have two ways to provision it

Given a network-wide forwarding state to provision, one can synthesize

way 1 the routing messages shown to the routers

way 2 the configurations run by the routers

Given a network-wide forwarding state

output to provision, one can synthesize

inputs the routing messages shown to the routers

functions the configurations run by the routers

Network control & programmability

through synthesis



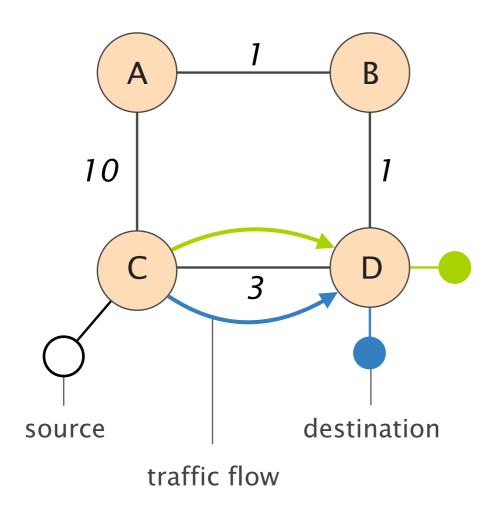
Network control & programmability

through synthesis

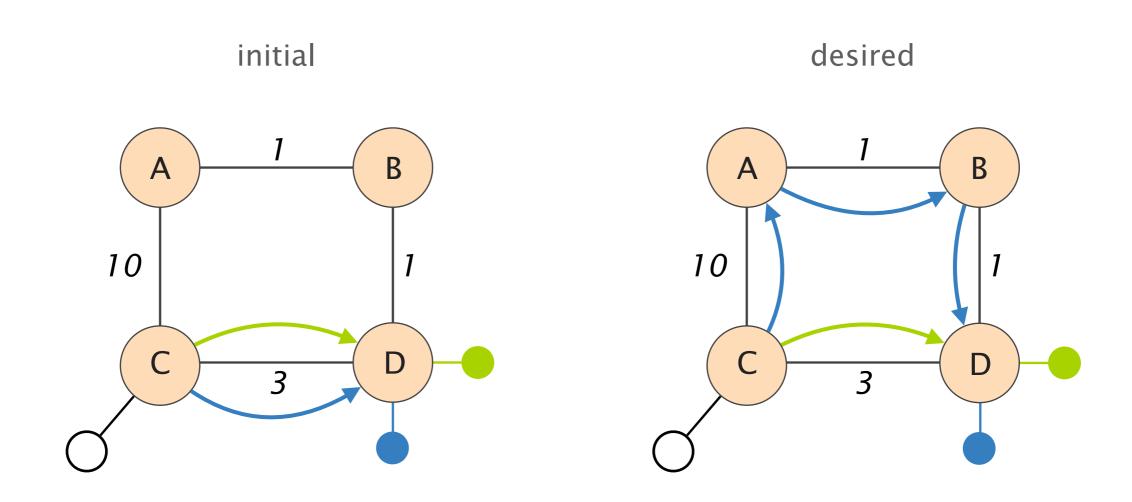


Part 1

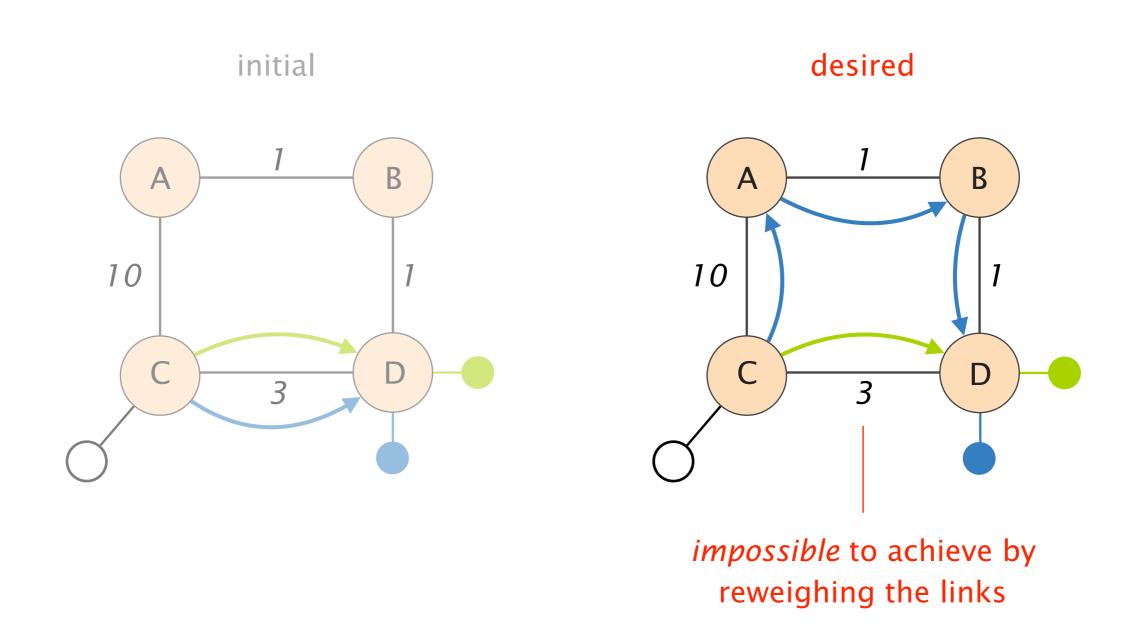
Consider this network where a source sends traffic to 2 destinations

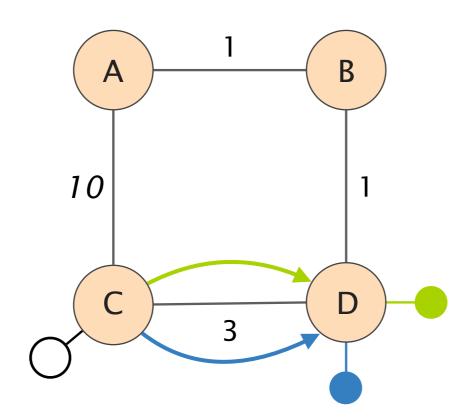


As congestion appears, the operator wants to shift away one flow from (C,D)

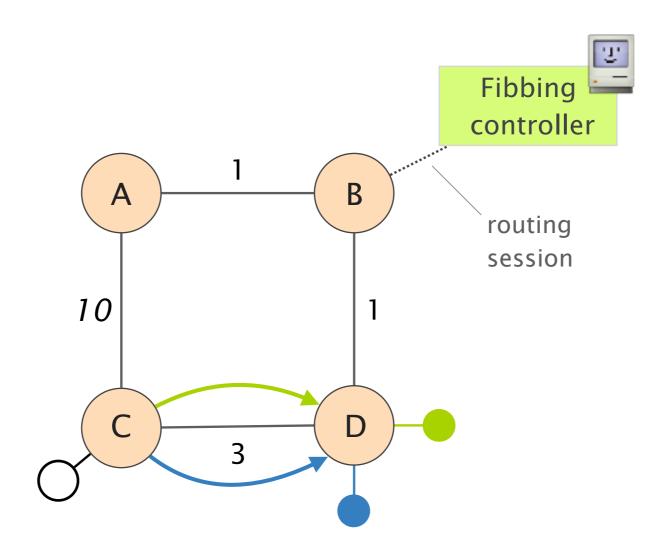


Moving only one flow is impossible though as both destinations are connected to D

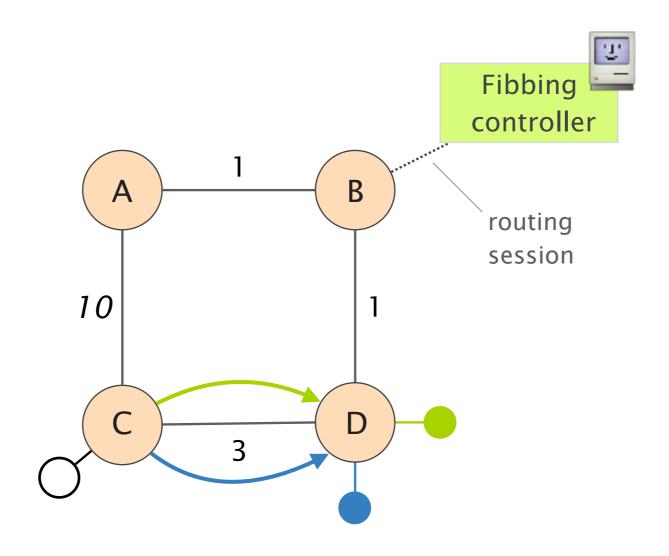


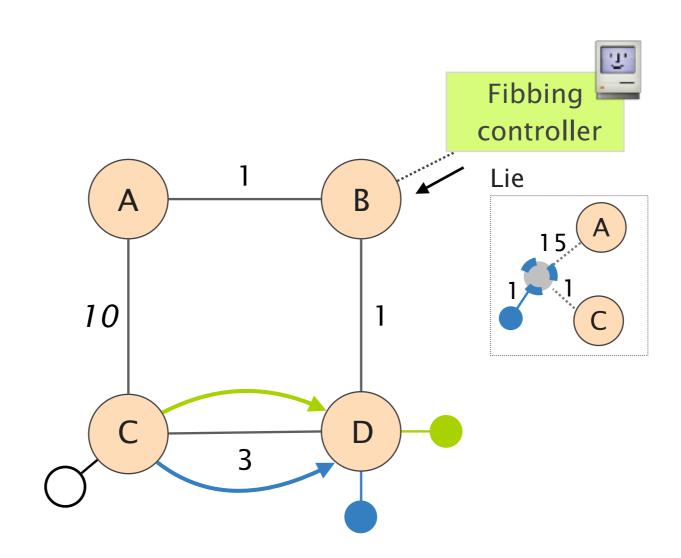


Let's lie to the routers

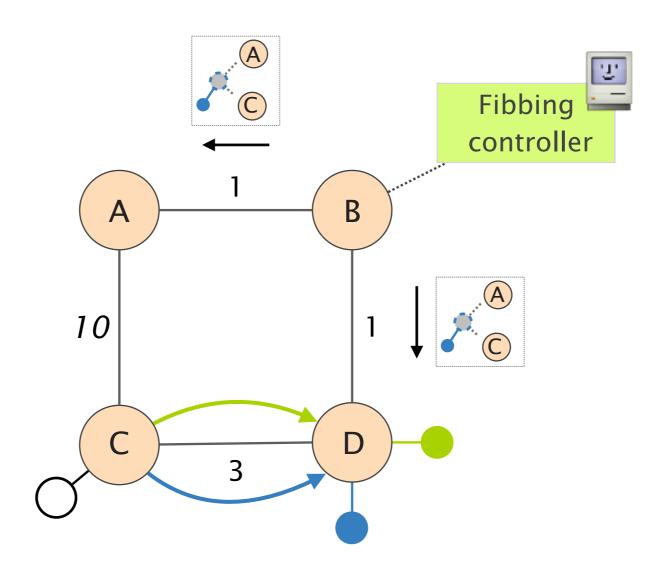


Let's lie to the routers, by injecting fake nodes, links and destinations

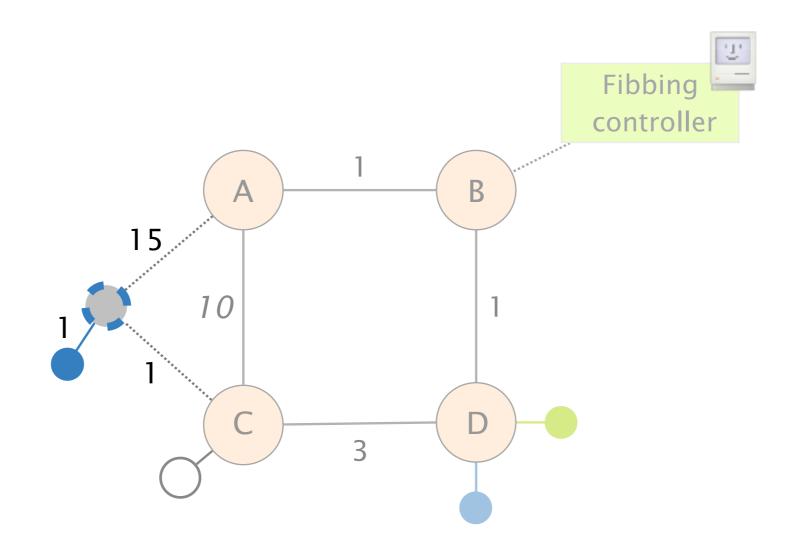




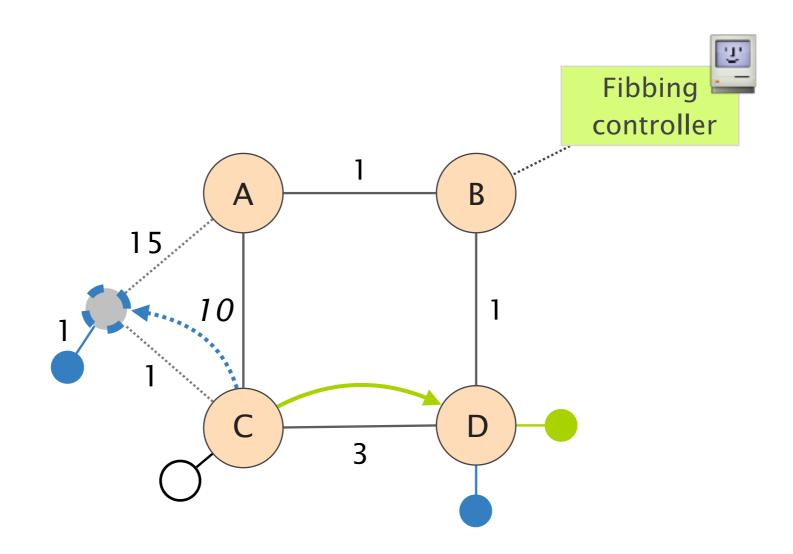
Lies are propagated network-wide by the routing protocol



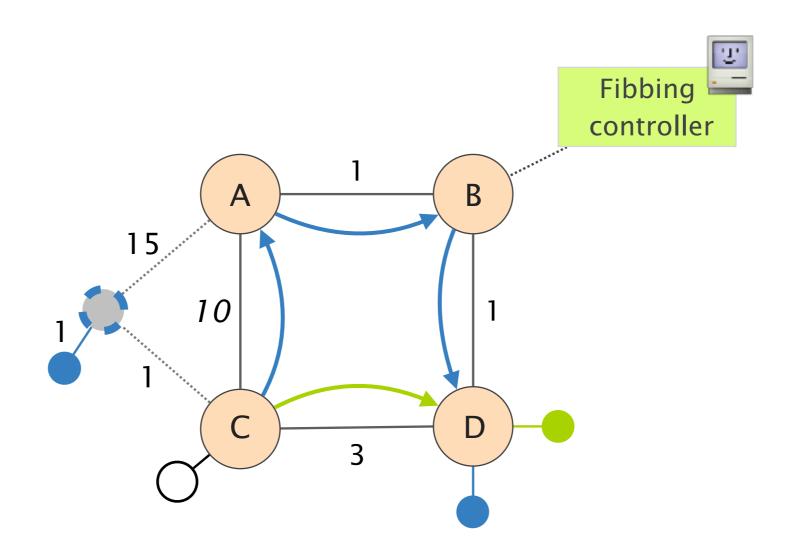
All routers compute their shortest-paths on the augmented topology



C prefers the virtual node (cost 2) to reach the blue destination...



As the virtual node does not really exist, actual traffic is physically sent to A



Synthesizing routing messages is powerful

Theorem

Fibbing can program any set of non-contradictory paths

Theorem

Fibbing can program

any set of non-contradictory paths

Theorem

Fibbing can program

any set of non-contradictory paths

any path is loop-free

(e.g., [s1, a, b, a, d] is not possible)

paths are consistent

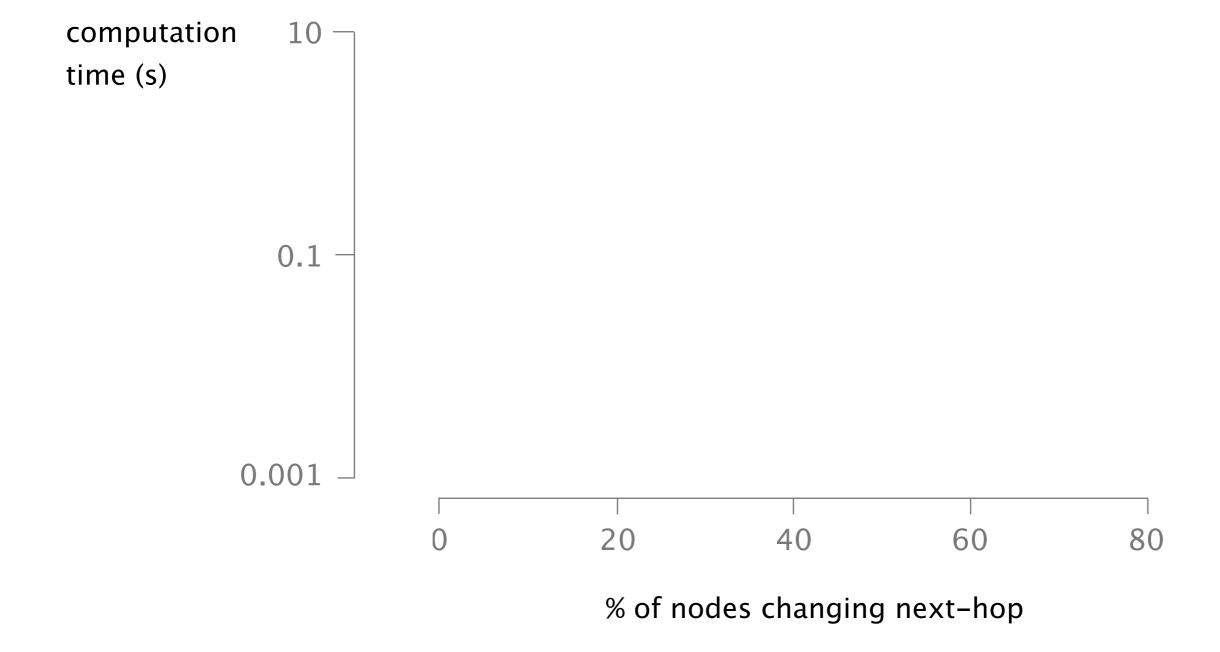
(e.g. [s1, a, b, d] and [s2, b, a, d] are inconsistent)

Synthesizing routing messages is fast and works in practice

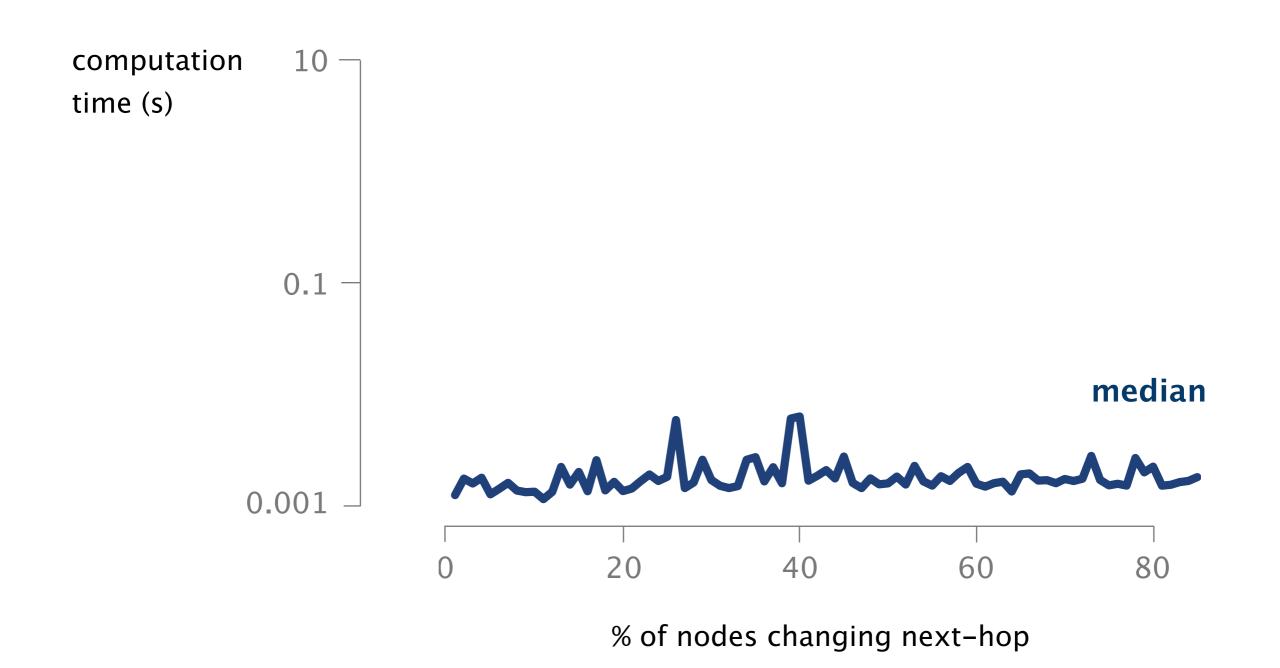
We developed efficient algorithms polynomial in the # of requirements

Compute and minimize topologies in ms independently of the size of the network

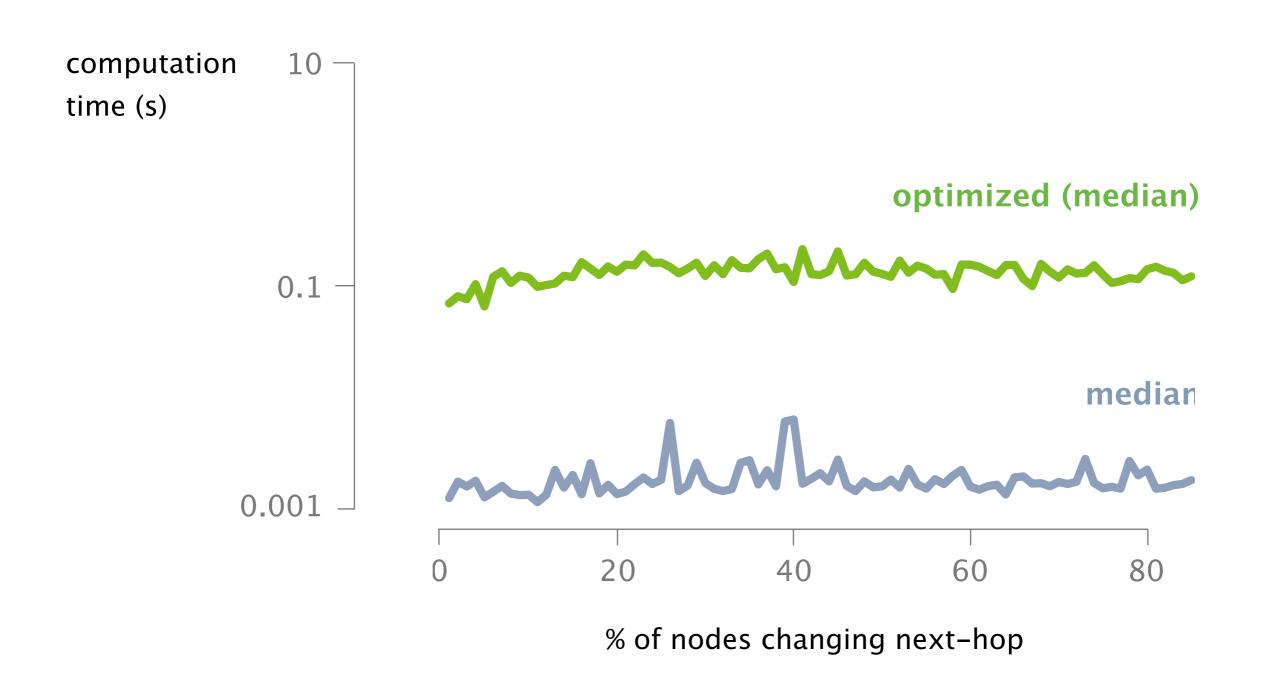
We tested them against real routers works on both Cisco and Juniper



Fibbing computes routing messages to inject in ~1 ms



Fibbing minimizes the # of routing messages to inject in ~100ms



Fibbing is fully implemented and works with real routers

Existing routers can easily sustain Fibbing-induced load, even with huge topologies

100 000

# fake	router	
nodes	memory (MB)	
1000	0.7	
5 000	6.8	
10 000	14.5	
50 000	76.0	

153

DRAM is cheap

Because it is entirely distributed, programming forwarding entries is fast

fake installation nodes time (s)

0.9

5 000 4.5

1000

10 000 8.9

50 000 44.7

100 000 89.50

894.50 μs/entry

Fibbing is limited though, among others by the configurations running on the routers

Works with a single protocol family

Dijkstra-based shortest-path routing

Can lead to loads of messages if the configuration is not adapted

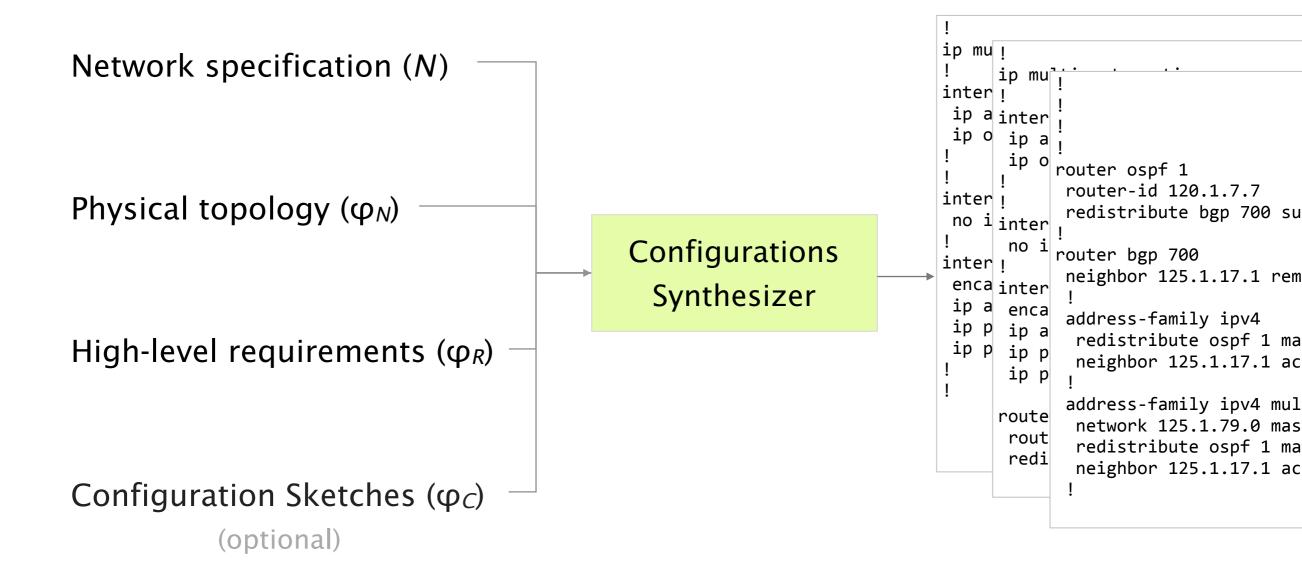
Suffers from reliability issues need to remove the lies upon failures

Network control & programmability

through synthesis



Part 2



Network specification (N)

Physical topology (ϕ_N)

High-level requirements (φ_R)

Configuration Sketches (φ_c)

Network specification (N)

Physical topology (ϕ_N)

High-level requirements (φ_R)

Configuration Sketches (φ_c)

A model of how the routers compute their forwarding state as a function of their configuration

Network specification (N)

Physical topology (ϕ_N)

High-level requirements (φ_R)

Configuration Sketches (ϕ_c)

An encoding of the physical topology

Network specification (N)

Physical topology (ϕ_N)

High-level requirements (φ_R)

Configuration Sketches (φ_c)

A set of constraints over the forwarding state produced by the synthesized configurations

Network specification (N)

Physical topology (ϕ_N)

High-level requirements (φ_R)

Configuration Sketches (ϕ_C) (optional)

A set of constraints on the content of the synthesized configurations

problem

Given N, φ_N , φ_R , φ_C

Generate a network-wide configuration C s.t. the ϕ_{N} , ϕ_{R} , and ϕ_{C} constraints are satisfied for the given network specification N

problem

Given N, φ_N , φ_R , φ_C

Generate a network-wide configuration C s.t. the φ_N , φ_R , and φ_C constraints are satisfied for the given network specification N

challenge

this is undecidable (in general)

problem

Given N, φ_N , φ_R , φ_C

Generate a network-wide configuration C s.t. the φ_N , φ_R , and φ_C constraints are satisfied for the given network specification N

challenge

this is undecidable (in general)

insights

domain-specific heuristics

Network Configuration synthesis: a booming field!

Out of high-level requirements, synthesize

Genesis [POPL'17]

static routes

doesn't support distributed routing protocols

Propane [SIGCOMM'16]

BGP configurations

PropaneAT [PLDI'17]

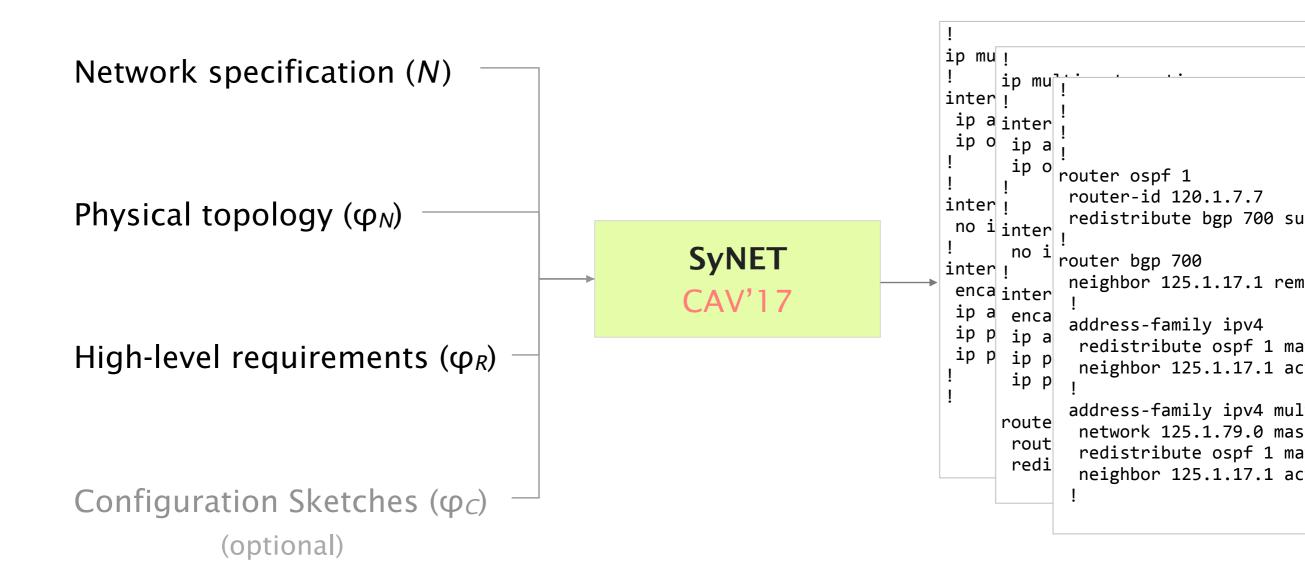
doesn't support other protocols

CPR [SOSP'17]

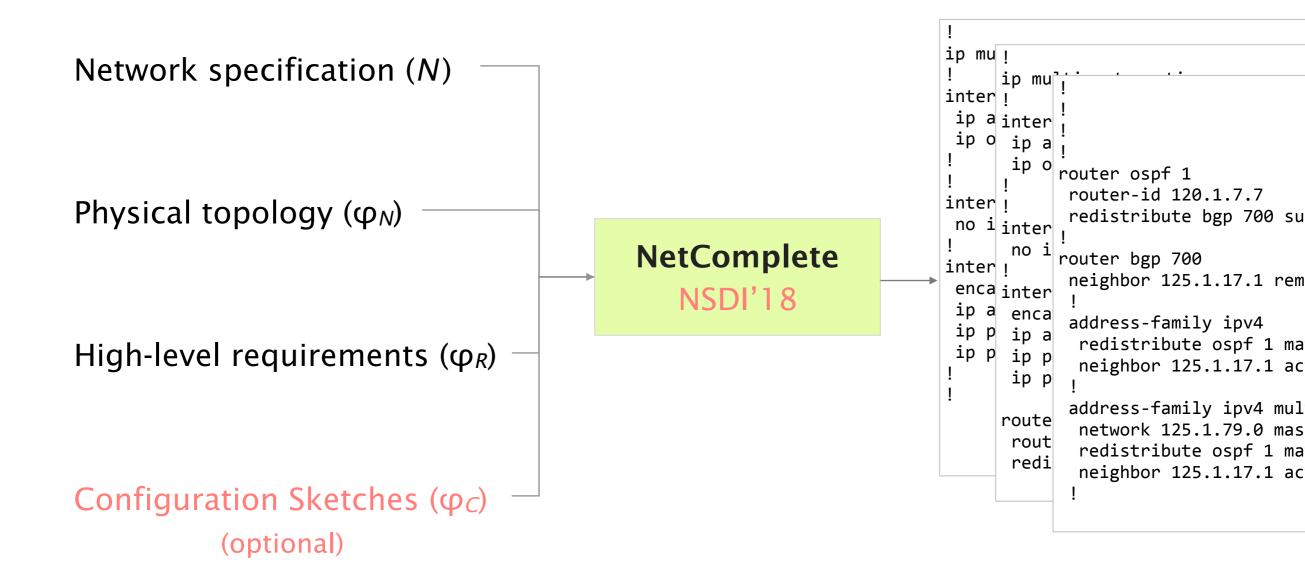
Minimal configuration repairs

partial support for BGP policies

Inputs



Inputs



By accepting partial configurations as inputs, NetComplete solves two problems of SyNET Problem #1
Interpretability

Existing synthesizers can produce configurations that differ widely from existing ones

Problem #1

Interpretability

Existing synthesizers can produce configurations

that differ widely from existing ones

NetComplete

Operators can control the synthesizer output

Problem #1

Interpretability

Existing synthesizers can produce configurations

that differ widely from existing ones

NetComplete

Operators can control the synthesizer output

Problem #2

Scalability

Large search space doesn't bode well with performance

Problem #1

Interpretability

Existing synthesizers can produce configurations that differ widely from existing ones

NetComplete

Operators can control the synthesizer output

Problem #2

Scalability

Large search space doesn't bode well with performance

NetComplete

Partial configurations reduce the search space

By accepting partial configurations as inputs, NetComplete solves two problems of SyNET

How? Using a "sketching language"

A configuration sketch is a configuration containing "holes" that have to be synthesized

```
interface TenGigabitEthernet1/1/1
  ip address ? ?
  ip ospf cost 10 < ? < 100
router ospf 100
router bgp 6500
  neighbor AS200 import route-map imp-p1
  neighbor AS200 export route-map exp-p1
ip community-list C1 permit ?
ip community-list C2 permit ?
```

```
route-map imp-p1 permit 10
  set ?
  set ?
route-map exp-p1 ? 10
  match community C2
route-map exp-p2 ? 20
  match community C1
...
```

The holes can identify specific attributes such as IP addresses, link costs or BGP local preferences

```
interface TenGigabitEthernet1/1/1
  ip address ? ?
  ip ospf cost 10 < ? < 100
router ospf 100
router bgp 6500
 neighbor AS200 import route-map imp-p1
  neighbor AS200 export route-map exp-p1
ip community-list C1 permit ?
ip community-list C2 permit ?
```

```
route-map imp-p1 permit 10
set ?
set ?
route-map exp-p1 ? 10
match community C2
route-map exp-p2 ? 20
match community C1
...
```

The holes can also identify entire pieces of the configuration

```
interface TenGigabitEthernet1/1/1
 ip address ? ?
 ip ospf cost 10 < ? < 100
router ospf 100
router bgp 6500
 neighbor AS200 import route-map imp-p1
 neighbor AS200 export route-map exp-p1
ip community-list C1 permit ?
ip community-list C2 permit ?
```

```
route-map imp-p1 permit 10
set ?
set ?
route-map exp-p1 ? 10
match community C2
route-map exp-p2 ? 20
match community C1
...
```

The sketching language also allow operators to specify constraints on the concrete values

```
interface TenGigabitEthernet1/1/1
 ip address ? ?
  ip ospf cost 10 < ? < 100
router ospf 100
router bgp 6500
 neighbor AS200 import route-map imp-p1
  neighbor AS200 export route-map exp-p1
ip community-list C1 permit ?
ip community-list C2 permit ?
```

```
route-map imp-p1 permit 10
set ?
set ?
route-map exp-p1 ? 10
match community C2
route-map exp-p2 ? 20
match community C1
...
```

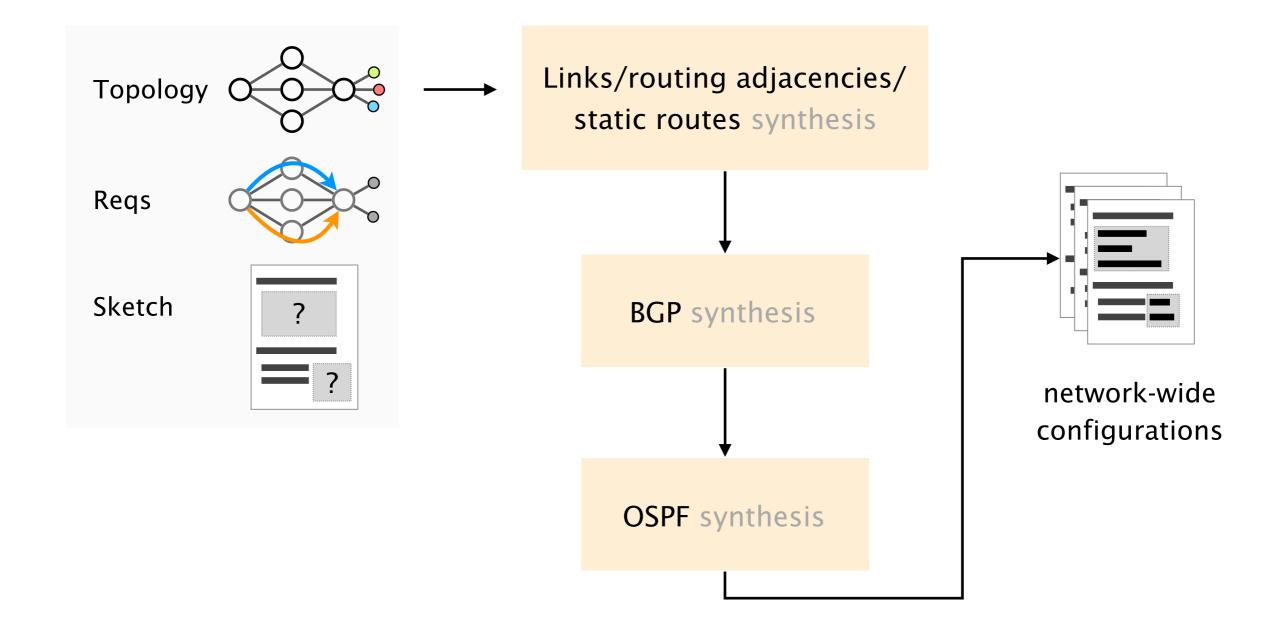
NetComplete "autocompletes" the holes such that the output configuration complies with the requirements

```
interface TenGigabitEthernet1/1/1
  ip address 10.0.0.1 255.255.255.254
  ip ospf cost 15
router ospf 100
  network 10.0.0.1 0.0.0.1 area 0.0.0.0
router bgp 6500
 neighbor AS200 import route-map imp-p1
  neighbor AS200 export route-map exp-p1
ip community-list C1 permit 6500:1
ip community-list C2 permit 6500:2
```

```
route-map imp-p1 permit 10
set community 6500:1
set local-pref 50
route-map exp-p1 permit 10
match community C2
route-map exp-p2 deny 20
match community C1
...
```

NetComplete

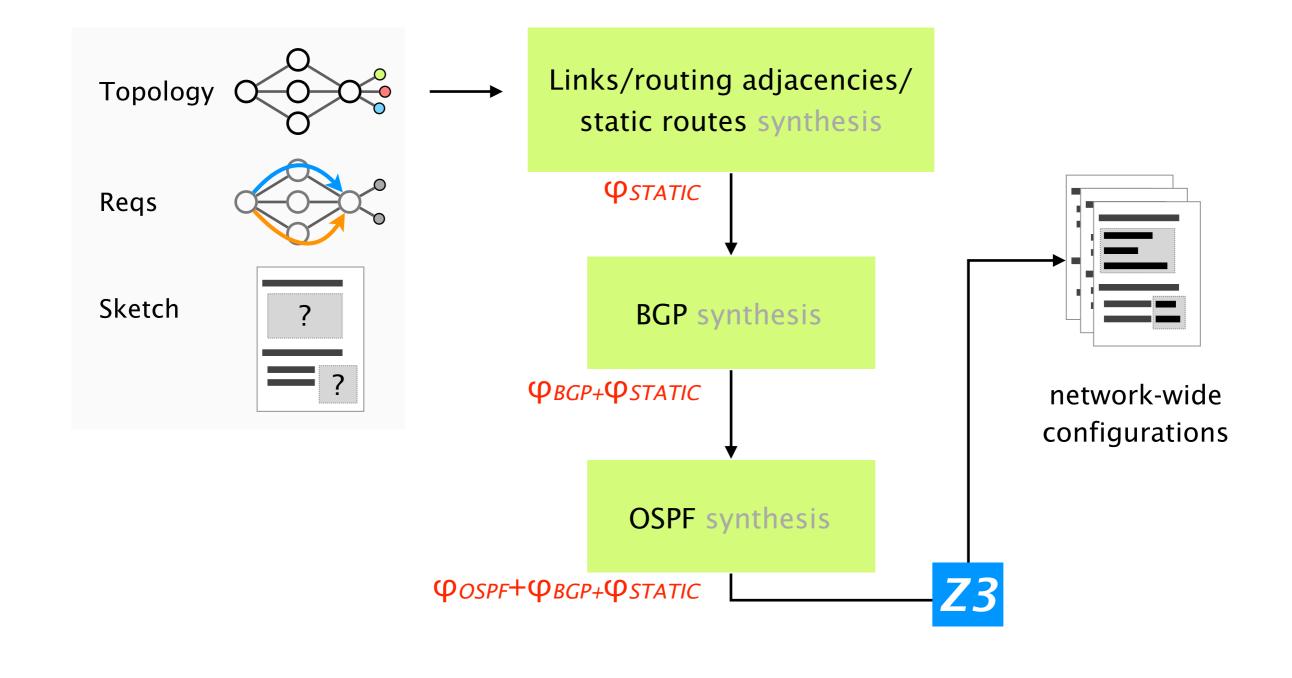
Outputs



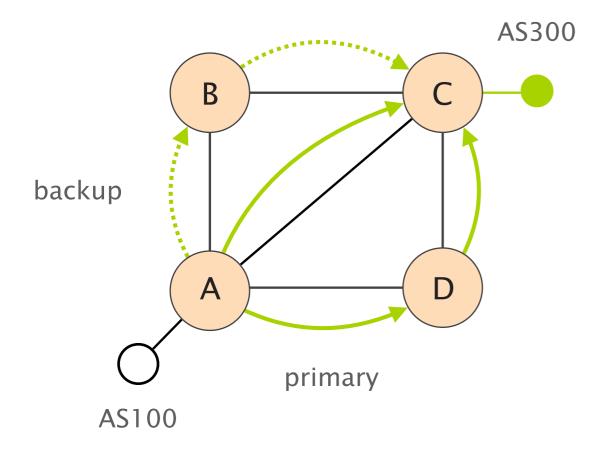
With respect to SyNET, NetComplete generates domain-specific SMT constraints for each protocol

NetComplete

Outputs

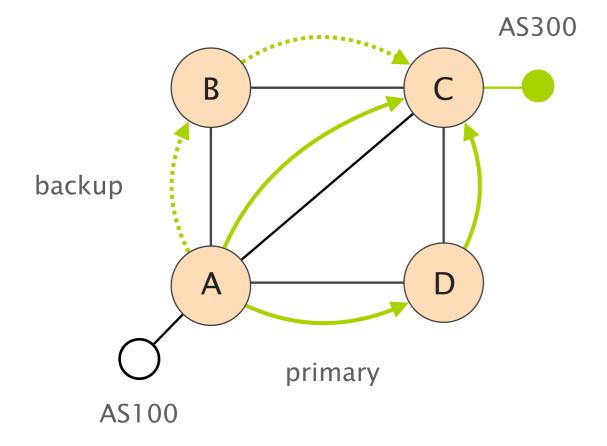


Let's consider the OSPF box



Requirement

$$(AS100 \rightarrow A \rightarrow C \rightarrow AS300 = AS100 \rightarrow A \rightarrow D \rightarrow C \rightarrow AS300) \gg AS100 \rightarrow A \rightarrow B \rightarrow C \rightarrow AS300$$



Requirement

$$(AS100 \rightarrow A \rightarrow C \rightarrow AS300 = AS100 \rightarrow A \rightarrow D \rightarrow C \rightarrow AS300) \gg AS100 \rightarrow A \rightarrow B \rightarrow C \rightarrow AS300$$

Naive OSPF encoding

$$Cost(A \to C) = Cost(A \to D \to C)$$

$$\land (Cost(A \to C) < Cost(A \to B \to C))$$

$$\land (\forall X \in Paths(AS100, AS300) \setminus S.$$

$$Cost(A \to B \to C) < Cost(X)), \text{ where}$$

$$S = \{A \to C, A \to D \to C, A \to B \to C\}$$

doesn't scale to large networks!

To scale, NetComplete leverages Counter-Example Guided Inductive Synthesis

To scale, NetComplete leverages

Counter-Example Guided Inductive Synthesis

An contemporary approach to synthesis where a solution is iteratively learned from counter-examples

While finding weights is hard, computing shortest-path is *easy*

CEGIS Consider a random subset 5 of them and

Part 1 synthesize the weights considering 5 only

Consider a random subset *S* of them and synthesize the weights considering *S* only

intuition

Fast as *S* is small compared to all paths, but can be wrong because we don't consider all paths

Consider a random subset *S* of them and synthesize the weights considering *S* only

CEGIS

Part 2

Check whether the weights found comply with the requirements over all paths

If so, return

If not, take a counter-example (a path) that violates the Req and add it to *S*

Repeat.

Consider a random subset *S* of them and synthesize the weights considering *S* only

Check whether the weights found comply with the requirements over all paths

intuition

Fast too

simple shortest-path computation

The entire procedure usually converges in few iterations—making it very fast in practice

Instead of considering *all* paths between *X* and *Y*

Consider a random subset *S* of them and synthesize the weights considering *S* only

Check whether the weights found comply with the requirements over all paths

If so, return

If not, take a counter-example (a path)

that violates the Req and add it to S

Repeat.

We fully implemented NetComplete and showed its practicality

Code

~10K lines of Python

SMT-LIB v2 and Z3

Input

OSPF, BGP, static routes

as partial and concrete configs

Output

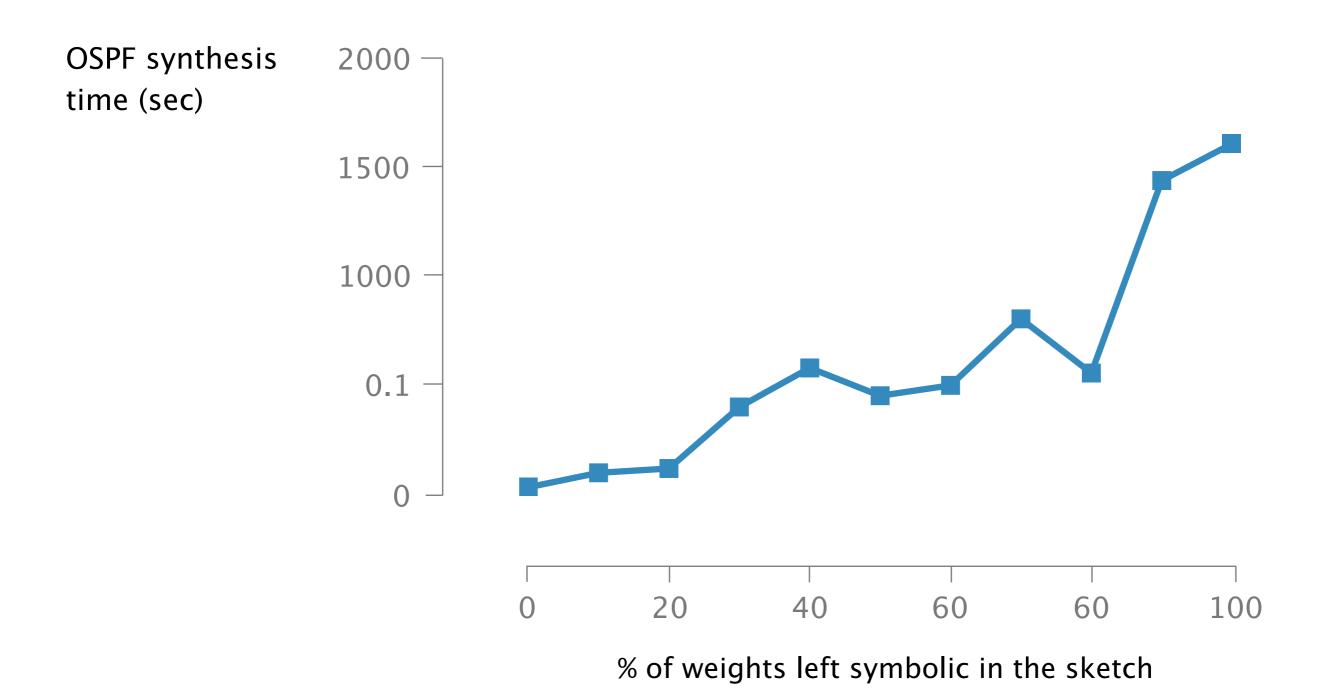
Cisco-compatible configurations

validated with actual Cisco routers

NetComplete can synthesize configurations for large networks in few minutes

	Network size	Reqs. type	-	of symbolic value in the sketch	
		16 in total	50%	100%	
OSPF synthesis time (sec) averaged over 5 topos	Medium 68—74 nodes	Simple	6s	6s	
		ECMP	6s	6s	
		Ordered	31s	43s	
	Large 145—197 nodes	Simple	14 s	14 s	
		ECMP	13s	14s	
		Ordered	249s	1155s	

NetComplete synthesis time increases as the sketch becomes more symbolic



NetComplete scales much better than SyNET

# routers	# protocols	SyNET	NetComplete
49	static	14 min	0.05s
	static, OSPF	5h 22min	2m 1s
	static, OSPF, BGP	timeout (>24h)	44m 2s
64	static	49 min	0.06s
	static, OSPF	21h 13min	2m 22s
	static, OSPF, BGP	timeout (>24h)	6h 6min

>600x speed-up **SyNET NetComplete** # routers # protocols 49 14 min 0.05s static static, OSPF 5h 22min 2m 1s 44m 2s static, OSPF, BGP timeout (>24h) 49 min 0.06s 64 static static, OSPF 21h 13min 2m 22s

static, OSPF, BGP

6h 6min

timeout (>24h)

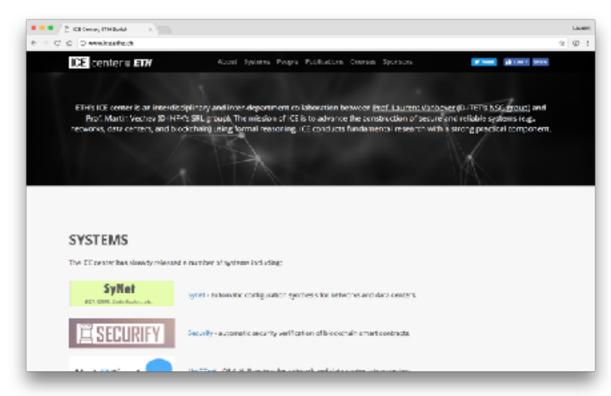
Network control & programmability

through synthesis



If you want to have more information about our research...





Networked Systems Group

nsg.ee.ethz.ch

ICE

ice.ethz.ch

Programming networks

Not your standard API





Laurent Vanbever

nsg.ee.ethz.ch

NII Shonan Meeting

Mon Feb 26 2018