Consistent Updates in Software-Defined Networks



Nate Foster Mark Reitblatt Cole Schlesinger Jennifer Rexford David Walker





Network Updates



Network Updates





At 12:47 AM PDT on April 21st, a network change was performed as part of our normal scaling activities...

During the change, one of the steps is to shift traffic off of one of the redundant routers...

The traffic shift was executed incorrectly and the traffic was routed onto the lower capacity redundant network.

This led to a "re-mirroring storm"...

During this re-mirroring storm, the volume of connection attempts was extremely high and nodes began to fail, resulting in more volumes left needing to re-mirror. This added more requests to the re-mirroring storm...

The trigger for this event was a **network** configuration change.



At 12:47 AM PDT on April 21st, a network change was performed as part of our normal scaling activities...

During the change, one of the steps is to shift traffic off of one of the redundant routers...

The traffic shift was executed incorrectly and the traffic was routed onto the lower capacity redundant network.

This led to a "re-mirroring storm"...

During this re-mirroring storm, the volume of connection attempts was extremely high and nodes began to fail, resulting in more volumes left needing to re-mirror. This added more requests to the re-mirroring storm...

The trigger for this event was a **network** configuration change.





Consensus Routing: The Internet as a Distributed System

John P. John" Ethan Katz-Bassett" Arvind Krishnamurthy" Thomas Anderson" Arun Venkataramani[®]

Abstract

Internet routing protocols (BGP, OSPF, RIP) have traditionally favored responsiveness over consistency. A router applies a received update immediately to its for-(a safety property), Internet routing has lost both.

To this end, we present consensus routing, a consistencyfirst approach that cleanly separates safety and liveness terdomain traffic engineering [1] because they have to using two logically distinct modes of packet delivery: a worry about its poorly understood side-effects. Perhaps stable mode where a route is adopted only after all depen-most tellingly, despite a decade of research investigating dent routers have agreed upon it, and a transient mode that the complex dynamics of interdomain routing, the goal heuristically forwards the small fraction of packets that of a simple, practical routing protocol that allows gen encounter failed links. Somewhat surprisingly, we find eral routing policies and achieves high availability has that consensus routing improves overall availability when remained elasive. used in conjunction with existing transient mode heuristics such as backup paths, deflections, or detorring. Ex-tr tabove goal. The key insight is to recognize consis-periments on the Internet's AS-level topology show that consensus routing eliminates nearly all transient discon-property and systematically separate the two design connectivity in BGP.

1 Introduction

propagating the update to other routers, including those understand. that potentially depend upon the outcome of the update. Consensus routing achieves this separation using two A thinks its route to a destination is via B but B disagrees,

*Dept. of Computer Science, Unix. of Washington, Seattle. †University of Massachusetts Amberet.

tens of seconds due to an inconsistent view among routers in a single autonomous system [44].

Our position is that the lack of consistency is at the root. warding table before propagating the update to other ity. First, protocol behavior is complex and unpredictable routers, including those that potentially depend upon the as routers by design operate upon inconsistent distributed outcome of the update. Responsiveness comes at the cost of routing loops and blackholes-a router A thinks its no indicator of when, if at all, the network converges to route to a destination is via B but B disagrees. By favor- a consistent state. Second, unpredictable behavior makes ing responsiveness (a liveness property) over consistency (a adety respects), Internet evolute has been been and the second between a second b as it is difficult to distinguish between expected behav-Our position is that consistent state in a distributed sys-tem makes its behavior more predictable and securable. fles innevation in the long term, e.g., network operators are reluctant to adopt protocol optimizations such as inmost tellingly, despite a docade of research investigating

Our primary contribution, consensus routing, achieves cems, thereby borrowing an old lesson from distributed system design. Consistency safety means that a router forwards a packet strictly along the path adopted by the Internet routing, especially interdomain routing, has tra-upstream routers unless the packet encounters a failed ditionally favored responsiveness, i.e., how quickly the link. Liveness means that the system reacts quickly to network reacts to changes, over consistency, i.e., ensuring failures or policy changes. Separating safety and liveness that packets traverse adopted routes. A router applies a re- improves end-to-end availability, and, perhaps more imceived update immediately to its forwarding table before portantly, makes system behavior simple to describe and

Responsiveness comes at the cost of availability: a router logically distinct modes of packet delivery: 1) A stable mode ensures that a route is adopted only after all depeneither because 1) B's old route to the destination is via dent routers have agreed upon a consistent view of global A, causing loops, or 2) B does not have a current route state. Every epoch, routers participate in a distributed to the destination, causing blackholes. BGP updates are snapshot and consensus protocol to determine whether or known to cause up to 30% packet-loss for two minutes or not updates are complete, i.e., they have been processed more after a routing change, even though usable physical by every router that depends on the update. The output of routes exist [26]. Further, transient loops account for 90% the consensus serves as an explicit indicator that routers of all packet loss according to a Sprint network study [16]. may adopt a consistent set of routes processed before the Even a recovering link can cause unavailability lasting snapshot. 2) A transient mode ensures high availability when a packet encounters a router that does not possess a stable route, either because the corresponding link failed

Consenses Rearing: The Internet as a Distributed System		
July P. July," Edus Kay, Boser," A rind Etchemarthy," "Roma Andersot"		
Abdrawd and descent days a localization on anong some		
The state is using another ideal is state in the state of the state o		
the approach the strategy spaces and processing are transmosting to the strategy str		
metricity is MEY anticity is MEY anticity is MEY anticity is methanism of equipment fluing. Conserve values are not a some fluing. A some fluing is a so		
Requirements once a first and of could be approximate of the part		
The additional sector of the sector s		

Ada P. Ada" Edus Kay Buset" A	rold Dishamethy' Roma Anderson'			
Anat Verlaturoman ²				
Abstract	ton of second class is an increasing time among material			
Interest managements in 2002 100% A MP test to some agents control quints consolitation to a fee weeking units taking arrangements to a fee weeking units taking arrangements to a fee weeking units taking arrangements to a fee of many large and the solution of the solution of the solution of the solution of the solution of the solution of many large and the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the fee of the solution of the solution of the solution of the solution of the solution of the solution of the fee of the solution of the s	The proton is the field of constrainty is a field on the set of t			
det numer fors gand specif, and a reactive such the functional forwards the cased fraction of partice the economic billed links. Incomellar sequencings, we that the consensus names properties and basic sensitive sequencies with anxiety transition and basic tion and a facing partic, infections or determing the partners on the forward's 100 best specifies also	of a single, provided manta present that allows per- end manip patients and relations light analysis for manual states. The present sector of the sector of the first sector period. The law region is a manufactories that allow a pair. The law region is a manufactories many as a well paper) and segmentations.			
menta mita deservata ante de seconomica de se	property and commutantly registers for two design one entrie, thereby history or of transit from doubleast system design. Consistency safety means for a restor			
Toward costing expectedly interdential scaling, for the distanticy format responsionses, i.e., fore quarkly the automb scale to changes, over constanting i.e., stanting dergalation terrors adoptions. A reader applies a to solved sphere terrors adoption to the terrors adoption at propagating the sphere to the terrors. And addressing these propagating the sphere to the terrors. And addressing these	Invanto content monty song tio path adquari to tio aprimen mone males the path adquare securities a black lab. Lessons mone there for quant-mone quality on halanss or pathor designs. Repressing unlive part lessons repress. And a databative single to describe and personity, subscripts behavior single to describe and minimum.			
but provide significant to strength the strength of the sphere. Responsions comes as the cost of a straightform A strength A strength to the strength of the straightform A strength A strength to the strength of the straightform A strength within the strength of the A strength of the strength of the A A strength burge, with 3 B strength of the strength of the A strength burge, with 3 B strength of the strength of the strength of the strength of the strength of the strength of the strength burger, the strength of the strength of the strength of the strength burger, the strength of the strength of the strength of the strength of the strength of the strength of the strength of the strength of the strength of the strength of the strength of the strength of the strength or strength of the strength of the strength of the strength	Concerns outing, affective like appendix relative linguish should be also of a plant design of the should be appendixed by a structure like a dispersive structure like and appendixes and appendixes of the should be appendixed by a structure like and appendixes			
from a momentary first can cannot associately being	maples. It is number and more ligh and diffy the spectra process of the second se			

ous Routing: The Internet as a Distributed System

R-BGP: Staying Connected In a Connected World

Nate Kushman Srikanth Kandula Dina Katabi Bruce M. Maggs nkushman@mit.edu kandula@mit.edu dk@mit.edu bmm@cs.emu.edu

ABSTRACT

Many studies show that, when Internet links go up or down, the dynamics of BGP may cause several minutes of packet loss. The loss occurs even when multiple paths between the sender and receiver domains exist, and is unwarranted given the high connectivity of the Internet.

Our objective is to ensure that Internet domains stay enected as long as the underlying network is connected. Our solution, R-BGP works by pre-computing a few strategically chosen failover paths. R-BGP provably guarantees that a domain will not become disconnected from any destination as long as it will have a policy-compliant path to to zero in both cases.

1 INTRODUCTION

It has long been known that during convergence, BGP, the Internet interdomain routing protocol, causes packet loss and transient disconnectivity. For example, Labovitz et al. show that a route change generates, on average, 30% packet loss for as long as two minutes [22]. Wang et al. report that a single routing event can produce hundreds of loss bursts, and some bursts may last for up to 20 seconds [35]. Both popular IP addresses with a lot of traffic as well as unpopular addresses suffer temporary discon-nectivity because of BGP dynamics [25, 31]. Furthermore, BGP causes much of the lasting transient failures that affect Internet usability; our recent paper [20] shows that half of VoIP outages occur within 15 minutes of a BGP update.

BGP often loses connectivity even when the underlying network continuously has a path between the sender and the receiver. Indeed, in the above studies, the underlying network continuously has such a path. The Internet topology is known for its high redundancy, even when considering only policy compliant interdomain paths [13, 36]. Hence, transient disconnectivity due to protocol dynamics policy compliant paths exist in the underlying network.

convergence times [9, 18, 29, 34]. Such approaches, however, are intrinsically limited by the size of the Internet and the complexity of the BGP protocol.

We take a fundamentally different approach. We focus on protecting data forwarding, Instead of trying to reduce the period of convergence, we isolate the data plane from any harmful effects that convergence might cause. Specifically, while waiting for BGP to converge to the preferred route, we set the data plane to forward packets on precomputed failover paths. Thus, packet forwarding can continue unaffected throughout convergence.

Our failover design addresses two important challenges: that destination after convergence. Surprisingly, this can (a) Ensuring Low Overhead: The size and connectivity be done using a few simple and practical modifications to of the Internet make a naive advertisement of alternate BGP, and, like BGP, requires announcing only one path per failover paths unscalable. Announcing multiple paths to neighbor. Simulations on the AS-level graph of the current each neighbor could lead to explosion of the routing state, Internet show that R-BGP reduces the number of domains and announcing even a single failover path per neighbor that see transient disconnectivity resulting from a link fail. could lead to excessive update traffic. Instead, in our deure from 22% for edge links and 14% for core links down sign, a domain announces only one failover path to one strategic neighbor.

> (b) Guaranteeing Continuous Connectivity: The real difficulty in using failover paths lies in ensuring connectivity while progressing from the failover state to the final converged state. Inconsistent state across Internet domains can cause forwarding loops, or lead domains to believe that no path to the destination exists even when such a path does exist. We address the consistency problem by annotating BGP updates with a small amount of information that prevents transient routing loops and ensures that forwarding is never updated based on inconsistent state.

Our solution, R-BGP (Resilient BGP), needs only a few simple and practical changes to current BGP. It precomputes a few strategically chosen failover paths and maintains enough state consistency across domains to ensure continuous path availability. R-BGP has these properties:

- · Two domains are provably never disconnected by the dynamics of interdomain BGP, as long as the underly-
- ing network has a policy compliant path between them. · Like BGP, R-BGP advertises only one path per neighbor, and thus the number of updates it produces is on a par with BGP.

We evaluate R-BGP using simulations over the actual is unwarranted. The objective of this work is to ensure that Internet AS topology. Our empirical results show that, when Internet domains are continuously connected as long as a link fails, R-BGP reduces the number of domains temporarily disconnected by the dynamics of interdomain BGP Past work in this area has focused purely on shrinking from 22% for edge links and 14% for core links down to zero in both cases. Even in the worst case when multiple



R-BGP: Staying Connected in a Connected World

Nam Kashman Silkardh Kandula Dina Kasali Bruze M. Magge Sashman Festure: Sandul Westure: dirite state in the Minister and

Graceful Network State Migrations

Saqib Raza, Member, IEEE, Yuanbo Zhu, and Chen-Nee Chuah, Senior Member, IEEE

Abstract—A significant fraction of network events (such as topology or route changes) and the resulting performance degra-dations stem. This paper introduces a general class of Grouefal Network State Mignation (CNSM) problems, where the goal is to discover the optimal sequence of performance degra-gration (CNSM) problems, where the goal is to discover the optimal sequence of performance degration (CNSM) problems, which typically involve migrating transition the network from its initial to a desired flast state while a network from its initial to a desired flast state while transition the network from its initial to a desired final state while minimizing the overall performance disruption. We investigate two specific GNSM problems: (a) Link Weight Researchment Scheduling (LWRS) studies the optimal ordering of link weight updates to migrate from an existing to a new link weight assign-ment, and (b) Link Maintenance Scheduling (LMS) looks at how to schedule link deactivations and subsequent reactivations for maintenance purposes. LWRS and LMS are both combinatorial optimization problems. We use dynamic programming to find the optimization when the problem size is small, and leverage Ants Colony Optimization to get near-optimal solutions for large problem sizes. Our simulation study reveals that lipidiciously problem incs. Our simulation study reveals that lipidiciously and approximation when theremover is generic and applies to similar problems with different operational contexts, underlying similar problems with different operational contexts, underlying network protocols or mechanisms, and performance metrics.

I. INTRODUCTION

The Internet has been an enabling technology for missioncritical applications and services such as Voice over IP, operators to re-optimize and reset link weights. In such a VPNs, e-commerce applications, and multimedia streaming. Such applications rely upon consistent Quality of Service (QoS) provisioning by Internet Service Providers (ISPs), with five-nines availability (99.999% uptime) becoming the norm rather than the exception. The end-to-end perceived QoS can potentially be affected due to the dynamic nature of networks. For instance, network topology may change due to transient router link outages or long-term network engineering. Furthermore, protocol configuration parameters may be altered to migrate from one setting to another. Ideally, QoS guarantees should persist across such dynamic conditions

Some of these dynamic changes are inadvertent e.g., ones pig. 1. Example Network due to faulty interfaces, router crashes, and accidental fiber cuts. However, other changes ensue from deliberate and premeditated actions of network operators (e.g., routine maintenance). A failure characterization study of an IP backbone weights. Suppose all links have capacity c, and traffic demands [18] observed that planned maintenance activities account for [18] observed that planned maintenance activities account for more than 20% of transient failures. Other studies [8] have demand between all other node pairs is 0. The link weights also observed the prevalence of such planned maintenance depicted in Fig. 1 are optimal for such a traffic matrix given activities. Premeditated network tasks also include network the objective of minimizing the maximum link utilization upgrade activities such as adding new routers or overhauling (MLU). Now suppose that the traffic demand between node link capacity. Another example of a premeditated network task is migrating an existing OSPF [20] or IS-IS [25]1 link weight assignment to a new assignment that has been optimized based that corresponds to all weights being 1 [10]. This means that on the most up-to-date traffic matrix estimates.

Most common intra-domain (IGP) protocols.

a series of atomic operations. Each of these operations may cause some performance disruption that is a function of the network's changed state. The GNSM problem is to discover the sequence of operations that progressively transition the network to the final state while minimizing the overall disruption. This paper looks at two specific GNSM problems, as described below.

A. Link Weight Reassignment Scheduling

Setting link weights is the primary tool used by network operators to control network load distribution and traffic engineering [9-11,22]. Link weights are optimized based on an estimate of the traffic matrix. They are usually not modified in response to short-term fluctuations in the traffic matrix. However, the estimated traffic matrix may change significantly over a longer period of time, prompting network setting to another. The sequence in which the link weights are changed determines the disruption to network traffic during this migration process.

We illustrate this with the help of a toy example. Fig. 1 gives a network with the arc labels representing KIP link pair (b, g) increases to 3c. Shortest path routing using Equal Cost Multi-Path (ECMP) yields a new optimal weight setting three links weights (w(c, e), w(c, f), and w(d, f)) have to be

²This paper is an extended version of our previous work (28).



Consenses Routing: The Internet as a Distributed System Ada P. Ada". Educ Kay, Baser". Arriad Extension() ". Thomas Anderson" Any Websterson ton of second data <text><text><text><text><text> Thermone sensing addresses like separative using new regardly denses which of particle addresses like the regardly denses which of particle address like the sensitive sensitive addresses addresses addresses addresses sensitive addresses particular the sensitive like sensitive like addresses addresses addresses addresses by the sensitive like addresses addresses addresses by the sensitive like addresses addresses addresses addresses addresses ad address addresses by the sensitive like address addresses addresses by the sensitive addresses ad address addresses addresses addresses addresses ad address addresses addresses

R-BGP: Staying Connected in a Connected World

Nate Kusterum Silkardt-Kandula Dira Katala Bruce M. Higge dashman/Frobula: Kanlala/Frobula: do/Frobula: Inter/Frobula-

and a ageneth We have

for example (second response), (ter brog in 10 mil The density of the second ----

Dynamic Route Computation Considered Harmful

Matthew Caesar University of Illinois at Urbana-Champaion caesar@cs.illinois.edu

Martin Casado Teemu Koponen Nicira Networks Inc. casado@nicira.com Nicira Networks Inc. koponen@nicira.com

> Scott Shenker University of California at

Berkeley

Jennifer Rexford Princeton University jrex@cs.princeton.edu shenker@cs.berkeley.edu

ABSTRACT

This paper advocates a different approach to reduce routing convergence-side-stepping the problem by avoiding it in the first place! Rather than recomputing paths after temporary topology changes, we argue for a separation of timescale between offline computation of multiple diverse paths and online spreading of load over these paths. We believe decoupling failure recovery from path computation leads to networks that are inherently more efficient, more scalable, and easier to manage

Categories and Subject Descriptors

C.2.1 [Computer-Communication Networks]: Network Architecture and Design

General Terms

Algorithms, Design

Keywords

Internet architecture, routing, convergence, protocola

1 Introduction

The traditional task assigned to routing is very clear: routing calculates paths based on the current view of the network topology. When the network changes, whether due to a transient failure¹ or a permanent change in the topology, routing recomputes these paths. In fact, the reactive nature of routing-dynamically computing paths in response to failures to ensure connectivity-is typically seen as crucial to the Internet's resilience. We disagree.

¹This article is an editorial note submitted to CCR. It has NOT been peer reviewed. Authors take full responsibility for this article's technical content. Comments can be posted through CCR Online.

tarough C.R. Oulane. "The term "Allare" is overly narrow, since planned link maintenance is a common source of short-term service interruption, but for convenience we will use the term failure for all temporary outages.

In fact, there are several problems with using routing to respond to failure. First, computing a new set of paths with a distributed routing algorithm can be slow, leading to noticeable outages. Attempts to speed up convergence often risk scalability (by increasing the frequency of message eachanges) or involve ad hoc modifications (such as routeflap damping and tuning MRAI timers). Moreover, this recomputation can be for naught, since equipment failures and planned maintenance can be extremely transientoften completing before routing has reconverged-leading to another recomputation once the link is restored. In addition, route recomputation involves a complicated distributed algorithm that: is hard to understand and debug; is often the source of bugs, vulnerabilities, and misconfigurations; can lead to "update storms" if not properly tuned; and can only scalably compute a limited range of routing options (e.g., it is difficult to compute various sets of disjoint paths with a scalable distributed algorithm). Lastly, in some cases (such as spanning tree and BGP), the recomputation of routes doesn't just fix broken paths, it also disrupts working ones. In this paper, we argue that greater path diversity can, and should, reduce our reliance on failure-driven (i.e., "real-time") recomputation of routes. We are motivated by two trends in network and system design that improve an application's ability to tolerate failures:

- · Multipath routing lets edge nodes (whether end hosts or ASes) circumvent failures by explicitly directing traffic over a different path.
- · Data centers often replicate at the application layer, allowing a load balancer to direct flows to another server after a (host or network) failure.

We believe these trends should change the end-point's relationship with the network from 'TII accept whatever besteffort service you give me" to "TI take action to improve my service". In the early days of the Internet, the emphasis was on end-points passively "adapting" to the current level

Graceful Network State Migrations Ingli Ram, Mondor, HEE, Yambo Jhu, and Chen New Y

- X (In the case of permutational starts, network questions have permutative values of the supposed of the starts of the superscale of the starts of the superscale of the starts of the superscale of the superscale of the superscale of the superscale starts are superscaled as a superscale start of the superscale starts are superscaled as a superscale start of the superscale starts are superscaled as a superscale start of the superscale starts are superscaled as a superscale start of the superscale starts are superscaled as a superscale start of the superscale starts are superscaled as a superscale start of the superscale starts are superscaled as a superscale start of the superscale starts are superscaled as a superscale start of the superscale starts are superscaled as a superscale start of the superscale starts are superscaled as a superscale start are superscaled as a superscale starts are superscaled as a superscale starts are superscaled as a superscale starts are superscaled as a superscale start are supersca
polition man. For alcoholine shall strand that pathoants	A Link Weight Recongregation Tribuilding
pice, for 1999 phila featured, 1 picel and apply 5 body profiles risk Micro parabolic control, address principation of solution, ad printmass merci.	Intig 14 wight to be private test and by select species is could access that declarate air tasks regiments (P-1,2), 1,54 wights on applicable level or or extension of the selection for an applicable selection.
	mediled is segment to dest non decisation in he wells
 Bernausserertung 	ners from to annel tells ners no deap
The lossest issues a scaling achieving the sinite- related againstance and activate work a "Nan one it 120%, a conserved againstance, and antimeteric messangs facts againstance etc quare conserver Quility of ferrings Quility processing the lossest flow-to-the-losses QUIV, with first some constantion (10 MMA against flowing the core in some constantion) (10 MMA against flowing the core	equificantly over a logar period of time, prompting selected operators to exploring and time list engines. In such a case, network operators and its arguest lists are weight writing a ander. The segment is that the list engine are designed alternations the designers is activate soft, during the segments process.
and the to compare the entropy of proceed (pri- on pressial) is alread for a fit dynamic secon of activatio. For instance, activate topology into change due to	222-2
Reference, present colleptence prospects may be deter	04550-20
have of free distance charges an institution to a rest.	Re 1 Reach Second
day to have provident, more couldn, and exciting they are from the court of the definition and pro-	To there it will be here if a surmer fig. (
and not allow of allowing spectra in g., matter mate-	give a second with the an initial representing KIP link
"Harden A many second and the P harden	wright. Suggest of both fare repetity 1, and with the second
not the 374 of ression failure. Other makes (6) have	densed between of other ands pairs in it. The list weights
she descript its produces of and planed maintenance	depicted in Fig. 1 are optimal for such a readly matrix gives
service. Property of the set of the service of the service of the set of the	he signifies of annexing the maximum link attitution
int oracle. Some country of a president proved and	and had been a fe Barber and parties one find
a seguring an outsing (1677 (25) as 25-26 (25) Talk weight	Con Main Pub (SCMP) path a new spinst weight price
and present to a new animation that has been applicated based on the same array of the larger and the same an	the compact is of neight long 1 (4). The same the
	The second





R-BGP: Staying Connected in a Connected World

Nate Kashman Sikardi Kandula Dina Katali Bruce M. Hegge dashman Protody Lashid Protody di Protody Interneting

ABUTRUET	comparison (A. H. M. M. Industryaha, Inc.
α value dere Ha, vien benent bein proposition de la desense et HOT was ressource autore de la desense et al constante autore con teles autore con teles autore con teles autore con teles autore de la desense de la constante autore de la desense de la constante de la c	and an extense of p based by the case of the bases and the company of the HCP protocol. We show a start of the HCP protocol of energy and the extension of the protocol point of energy to the start of energy to the protocol point of the energy for the start of energy to the protocol point of the start of the start of the start of energy to the protocol point of the start of t
Assume that the set of the H and the processing approximation of the the set of the set	integrations and the particular particular term and the state of the
BORCOLUM Ing her land he brig stragges, IO in methods strag providence pairs instant decementary. For early, chiefe in the chief and any providence, it array, 19	B. Concentration of antipartic for association for the solution burgle is using the energy burgles for its researching connectivity of the programming from the Marine International Con- sequences and the solution of the solution of the solution on the solution programming and the solution of the solution and the solution programming and the solution of the solution public is the formation output research of the solution of the public is the formation output research of the solution of the solution public is the formation output research of the solution of the solution of the solution output the solution of the solution output the solution output to the so

In the trap of the transmission of the trap of the transmission o

Dynamic Route Computation Considered Harmful

Matthew Canzar Harrin Casado Terru, Kopuren nereta printen au Istory Menore Io. Kora Approxim Kala Dina Ali Alexinia Filosofta I. Sonto Filosofta Arostan Universi Product printen Ali product printen Ali

HOTEACT approx diverse is filter agreed in white unity approx diverse is publicly to setting to the origin of their is in sources, and of the setting to the setting to the setting to the setting we give requires in the public to the setting is gradied of the init the public to the setting is gradied of the setting to the setting is gradied of the setting is gradied of the setting to the setting is gradied to the setting to the setting is gradied to the setting to the set in the setting is gradient to the setting to the set in the setting is gradied to the setting to the set in the set into is collected to the set into point the set into the set is collected to the set into point the set into the set into is collected to the set into point the set into the set into the set is collected to the set into point the set into the set into the set is collected to the set into point the set into point the set into th	It is its does not served architect with same contrast to separat to billion. This is, sequepting a space of a spin- er part of the second sequences of a spine with a spin second sequences of a spine second sequences of the spin second second second second second sequences and the spin second second second second sequences and the spin second
where the bard is the control rate of the control rate. These is summarized theory, whether the to a same black of a spranger distribution of the state of the s	 a highly science bidden highly science big days do highly do not by down and by the science big days down and big days of the science big down and big days of the science big down and big days down and



Graceful Network State Migrations hup has, itedes, EEE, Tuele Jin, and Caro Yan Cash, Jean Horize, EEE

	In the case of parametrized with a sensor has questern been parameters to the the sensor of the sensor parameters are a sensor question of the sensor of the
anatolis and rough for advanta- order or effect applied of reach- tic and the second second second second data fragments a second second second data fragments and private second second metadates, and private second	A. Lind Wayle throughout Muniking Being Die wagke is its primes and and by select operate is somet annual bar distribution and table regimening (3-1).25. Link weights per optimized bard on a select of the solution space. New per second bard
1. Inclusion cross	mellini is segme to don'ten formation is its selfs name. Rearry, to animal selfs name my deep

These of the laptice index () is the interval $(1, \dots, n)$, is the probability of the index () is the ind

"Not street in store \$1" street.

America 6 angle regime of each of the second of the second second of the second of the second second of the second of the second of the second of the second second of the second

Software Abstractions



Software Abstractions



By designing the right software abstractions, we can solve the network update problem once and for all!

Software Abstractions



By designing the right software abstractions, we can solve the network update problem once and for all!



Security Policy

Src	Traffic	Action
	Web	Allow
	Non-web	Drop
	Any	Allow



Security Policy

Src	Traffic	Action
	Web	Allow
	Non-web	Drop
	Any	Allow

Configuration A

Process black-hat traffic on F1

Process white-hat traffic on {F2,F3}



Security Policy

Src	Traffic	Action
	Web	Allow
	Non-web	Drop
	Any	Allow

Configuration A

Process black-hat traffic on F1

Process white-hat traffic on {F2,F3}



Security Policy

Src	Traffic	Action
	Web	Allow
	Non-web	Drop
	Any	Allow

Configuration A

Process black-hat traffic on F1

Process white-hat traffic on {F2,F3}



Configuration B

Process black-hat traffic on {F1,F2}

Process white-hat traffic on F3

Abstractions for Network Update

Challenge

- The network is a distributed system
- Can only update one element at a time

Our Approach

• Provide programmers with constructs for updating the entire network at once

update(config, topo)

- Design semantics to ensure "reasonable" behavior
- Engineer efficient implementation mechanisms
 - Compiler constructs low-level update protocols
 - Automatically applies optimizations



Consistent Updates in Action



Main Function topo = Topo(...) update(configA, topo) ...wait for traffic load to shift... update(configB, topo)

Security Policy

Src	Traffic	Action
	Web	Allow
	Non-web	Drop
	Any	Allow



Semantics of Network Updates

Atomic Updates

- Seem sensible...
- but costly to implement
- and difficult to reason about effects on packets already in-flight



Semantics of Network Updates

Atomic Updates

- Seem sensible...
- but costly to implement
- and difficult to reason about effects on packets already in-flight

Per-Packet Consistent Updates

Every packet processed with old or new configuration, but not a mixture of the two



Implementation Mechanisms

Two-phase commit

- Construct versioned internal and edge configurations
- Phase 1: Install internal configuration
- Phase 2: Install edge configuration

Pure Extension

• Update strictly adds paths

Pure Retraction

Update strictly removes paths

Slice Update

• Update affects a small number of switches



(Ask me for a demo!)

(Ask me for a demo!)



(Ask me for a demo!)







Theorem

An update $u C_1$ to C_2 is per-packet consistent if and only if it preserves all properties satisfied by C_1 and C_2 .



Corollary



Corollary



Corollary



Corollary



Corollary

To verify that a property is invariant, simply check that the old and new configurations satisfy it

Properties

- Connectivity
- Loop freedom
- Blackhole freedom
- Access control
- Waypointing
- Totality



Per-Flow Consistency

Use Cases

- Load balancer
- Flow affinity
- In-order delivery

Per-flow consistent updates

Every *set* of related packets processed with old or new configuration, but not a mixture of the two.

Implementation mechanisms

- Need to identify active flows
- Rules with soft timeouts
- DevoFlow wildcard cloning
- End-host feedback



Ongoing Work

Other abstractions

- Loop-freedom
- Affinity preserving

Update Synthesis

- Programmer specifies an invariant
- Compiler constructs an update protocol

Enhanced fault tolerance

- Rapid response when failures occur
- Compiler "hardens" configurations
- Pre-loads backup policy

Leverage end hosts

Help identify active flows

Thank You!

Collaborators

Shrutarshi Basu (Cornell) Mike Freedman (Princeton) Rob Harrison (West Point) Chris Monsanto (Princeton) **Mark Reitblatt** (Cornell) Gün Sirer (Cornell) **Cole Schlesinger** (Princeton) Alec Story (Cornell) **Jen Rexford** (Princeton) **Dave Walker** (Princeton)

Funding





http://frenetic-lang.org