



ARISTA

Routing in Dense Topologies What's All the Fuss?

New Tools for Building Highly Scalable IP/MPLS Networks

Chris Martin, Tony Li
Arista Networks
October 2018

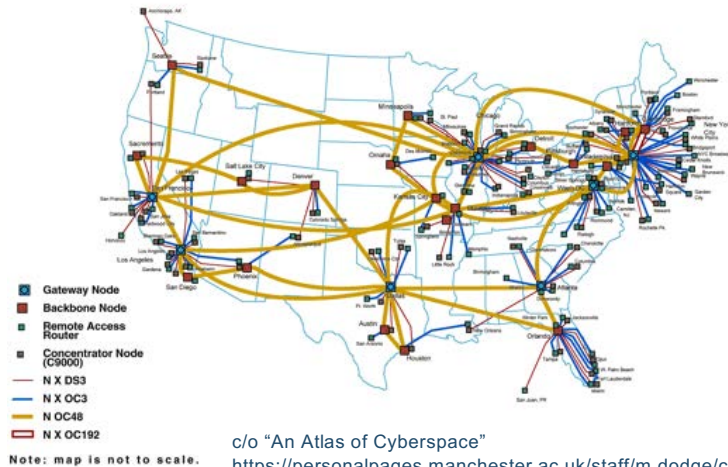
NANOG 74
Vancouver, Canada

Disclaimer

- This presentation contains the opinions of the authors only

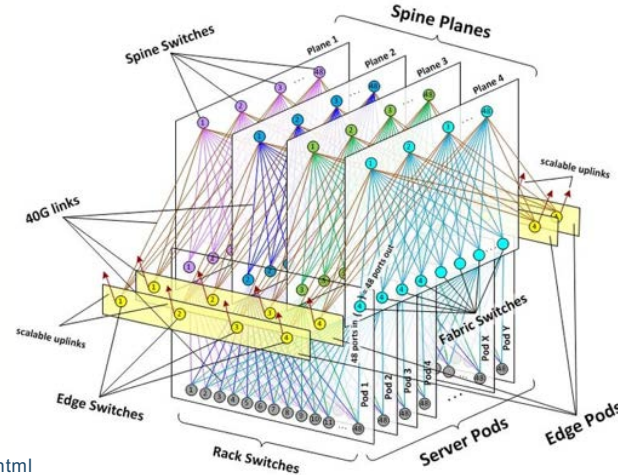
Introduction

- Rise of the Cloud has changed the architecture, design and scaling requirements of IP and MPLS networks
 - From flat, partial mesh topologies to hierarchical 3-D Clos networks



c/o "An Atlas of Cyberspace"

https://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/atlas/more_isp_maps.html



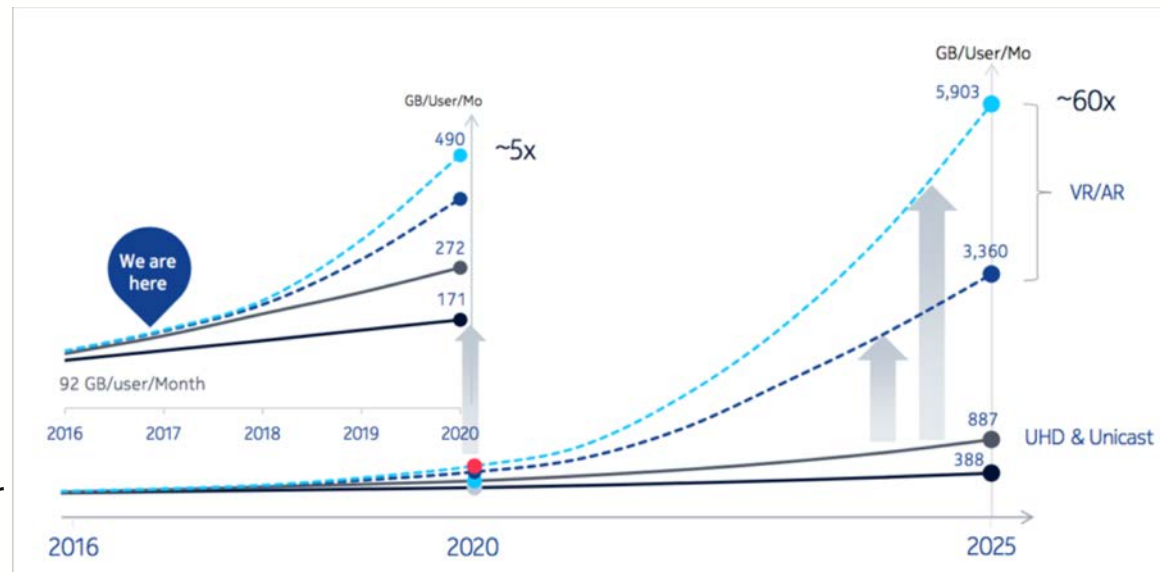
c/o Facebook Engineering
"http://code.facebook.com"

- IGP scaling limitations led to BGP adoption as de facto protocol for DC
 - However, BGP "dump truck" mentality overloads the semantics of the protocol
 - Loss of topology detail reduces value of IGP-based forwarding mechanisms (ie LFA)

How Did We Get Here?

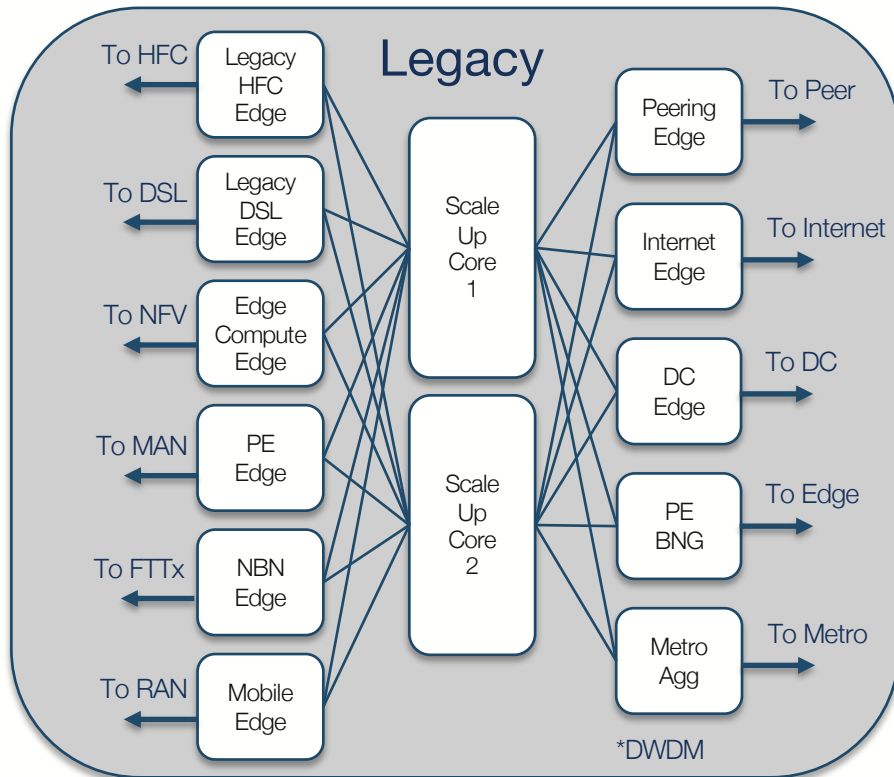
- Traditional WANs designed as sparse, uniplanar, polygonal lattices
 - Cost of long-haul links, router ports, and concerns about IGP scale
 - Routers vertically scaled (and integrated), allow for single control and management point
 - RSVP-TE scaling concerns due to soft, dense midpoint state and $O(n^2)$ tunnel scale
- Same design pattern in use for 20+ years

Unsustainable growth
using current network
design patterns!



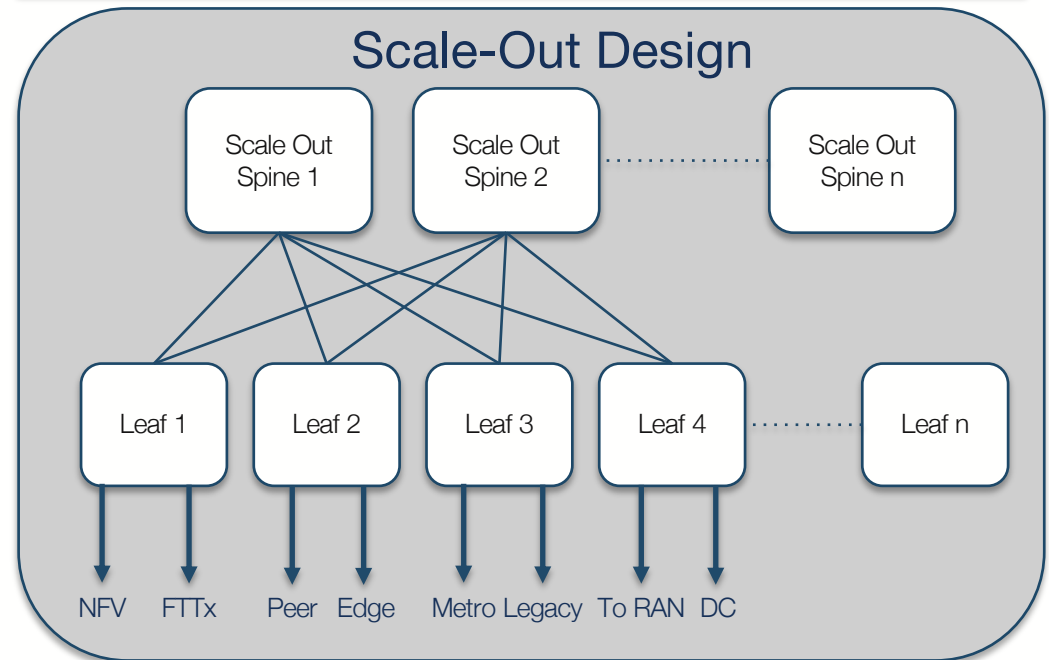
c/o Marcus Weldon, "The Future X Network", Nokia Bell Labs

Scale Up Vs Scale Out



Site Capacity Limited By Max Scale of one Core Router
 Difficult to take Core/Edge Routers Out of Service for Maintenance
 Single Purpose Edge Routers Increases CAPEX and Core Port Count

Scale-Out Isn't A Radical Change From Existing Carrier Topology Or Traffic Flows. It Is More Efficient.

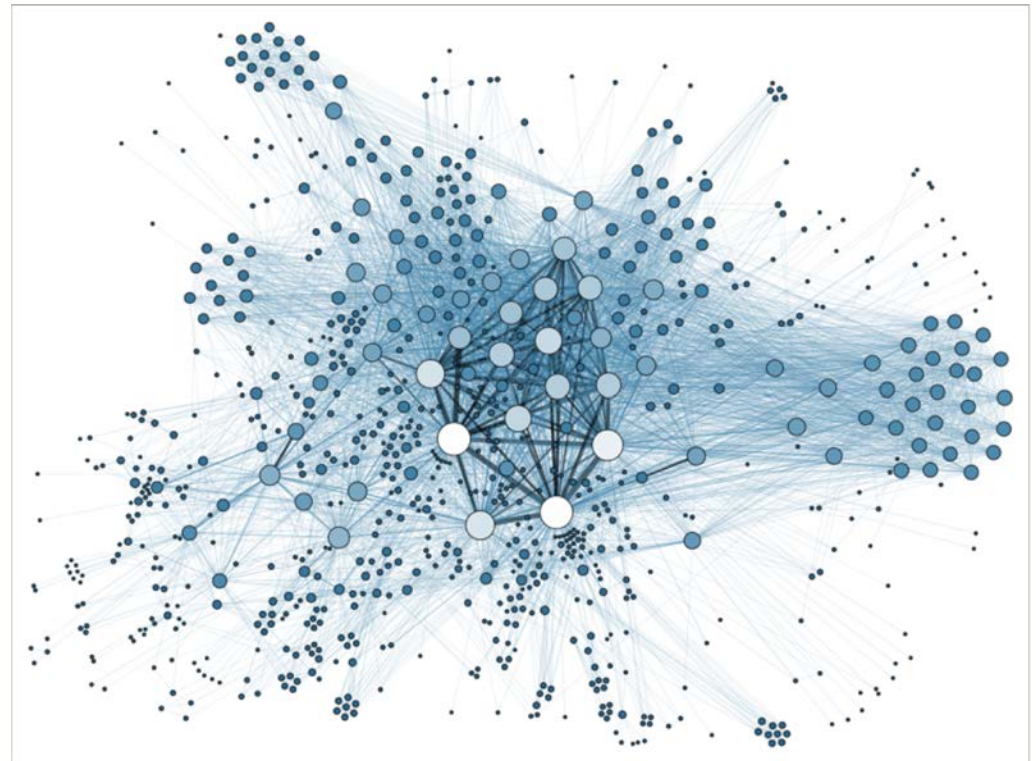


Elastic Site Capacity – Add More Spines or Leaf Nodes
 Simplified Operation with Higher Availability
 Scale Out Reduces CAPEX, Converges Edges, Provides Optionality

Challenges with Current Approach

- IGP scaling issues have (largely) led to widespread use of BGP in DCs
- BGP adoption in MSDC expanding
 - BGP app dev for policy control via tooling (ie, BIRD, Quagga, GoBGP)
 - Simple configuration if automated and scripted
 - Known scale, symmetric topology makes it almost trivial
 - ECMP L/S design reduces convergence wavefront to Clos stage diameter
- However, BGP can't really be used as an IGP in an arbitrary topology without an almost impossible amount of configuration
 - Unscriptable with irregular topologies
 - One router per AS = RIPv4

Need a solution to graphs like these, where the circular nodes may themselves be comprised of dense, bipartite graphs



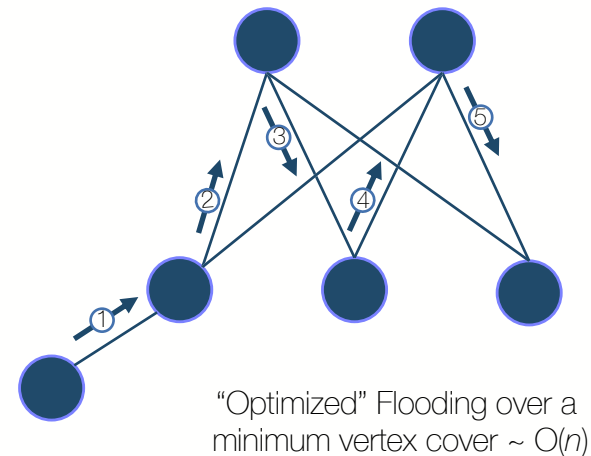
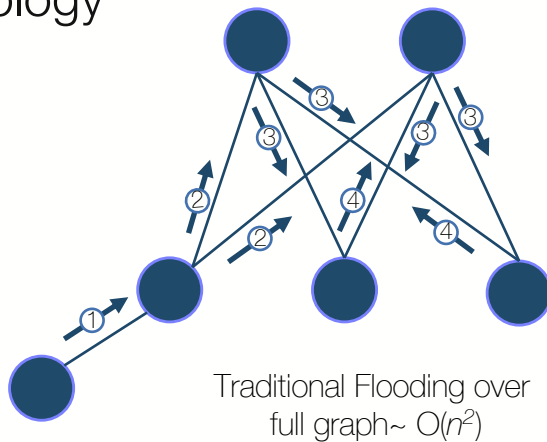
Brief History of IGP Scaling “Rules”


- John Moy OSPF Scaling Recommendations
 - 50 routers per area...
- Dave Katz – IS-IS OSPF Comparative Anatomy*
 - “80,000 LSA – refresh every 23 msec”
 - “ $O(n^4)$ information flooded during node failure”
 - “Don’t put zillions of routes in your IGP”
 - “Flooding load is the only real consideration”
- Unnecessary, redundant flooding is biggest scale inhibitor in IGPs
 - Especially IS-IS, where refresh interval is large (18.7 hours)

IGP Flooding Example

*animated

- IGP flooding is opportunistic and complete – flood everywhere while maintaining transmission lists to prevent endless reflooding, w/split horizon
- In dense, bipartite graphs, the amount of information flooded overwhelms the control plane at scale, with no solution to date other than avoidance
- Goal should be to reduce flooding to a minimal (not nec. optimal) flooding topology





IGP Dynamic Flooding and Area Abstraction/Hierarchy

draft-li-dynamic-flooding
draft-li-isis-area-abstraction
draft-li-isis-area-hierarchy

Importance of Link State Protocols in Modern Networks

- Information regarding the behavior and characteristic state of links in the network is easily conveyed in the IGP, which can later be used for critical forwarding plane operations
- Next generation multicast (BIER) and TE (both with Segment Routing and RSVP) benefit from a LS IGP
 - In the absence of a controller and detailed topology discovery, it is the only way to do Segment Routing, RSVP TE, and BIER
 - TI-LFA is critical for ensuring resilience without RSVP-TE FRR (which also requires an LS IGP)
- Importantly, the ability to extend detailed topology information as far across the network as possible alleviates the need for various hacks that aim to work around the loss of information at IGP area/level/process boundaries
 - Traditionally a challenge, due to IGP scaling limits (which reduce to flooding concerns)
 - Recall challenges around interarea link/node protection, inter-AS TE, inter-domain everything

What is draft-li-dynamic-flooding All About?

- In a dense topology, the flooding algorithm that is the heart of conventional link state routing protocols causes a great deal of redundant messaging.
 - This is exacerbated by scale.
- While the protocol can survive this combination, the redundant messaging is unnecessary overhead and delays convergence.
- Thus, the problem is to provide routing in dense, scalable topologies with rapid convergence.
- For this, we need a *flooding topology* that is a subset of the forwarding topology

Requirements for Dynamic Flooding

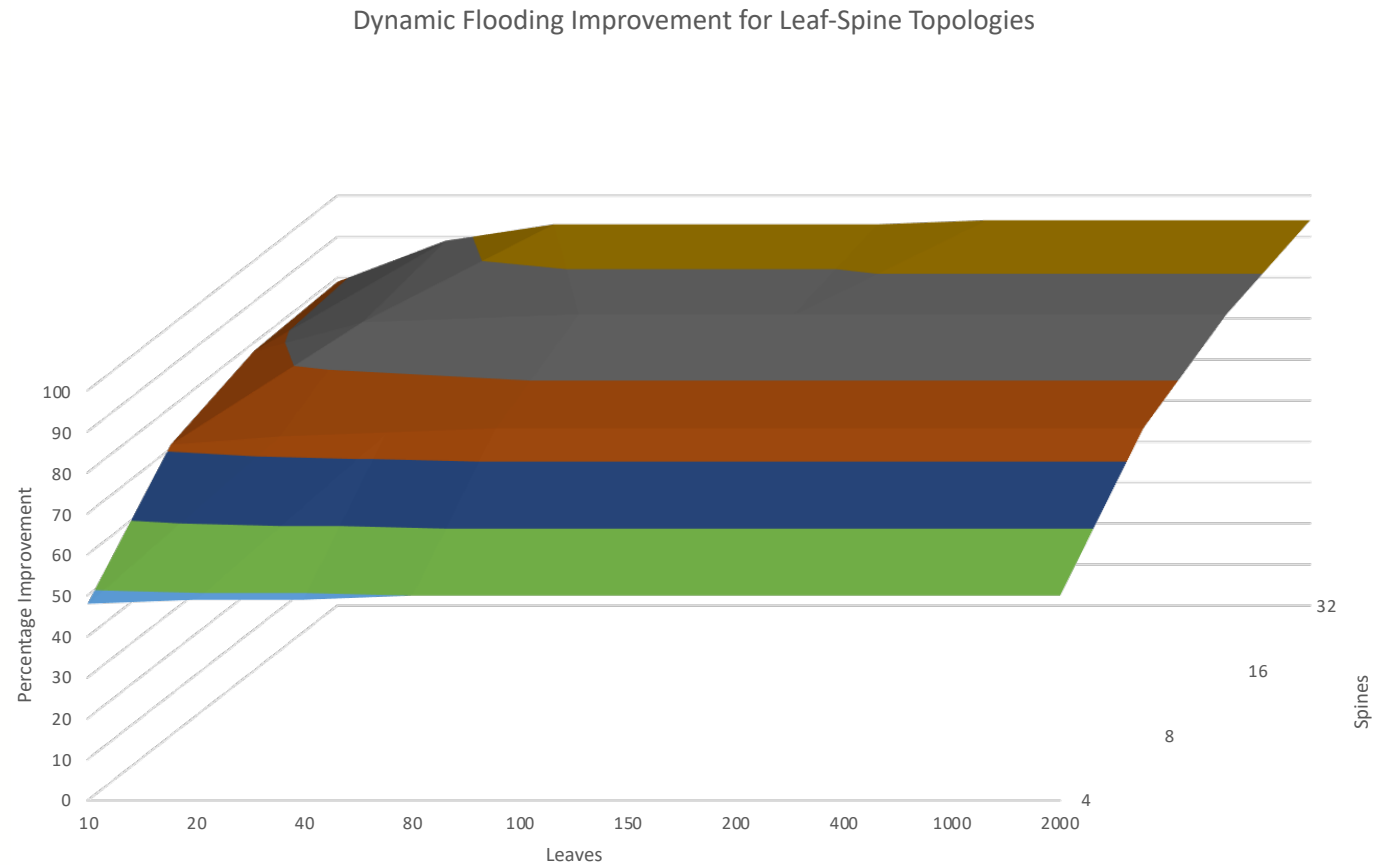
- Requirement 1: Provide a dynamic routing solution. Reachability must be restored after any topology change.
- Requirement 2: Provide a significant improvement in convergence.
- Requirement 3: The solution must address a variety of dense topologies.
 - Just addressing a complete bipartite topology such as $K_{5,8}$ is insufficient.
 - Multi-stage Clos topologies (and slight variants) must also be addressed.
 - Addressing complete graphs is a good demonstration of generality.
- Requirement 4: There must be no single point of failure. The loss of any link or node should not unduly hinder convergence.
- Requirement 5: Dense topologies are subgraphs of much larger topologies. Operational efficiency requires that the dense subgraph not operate in a radically different manner than the remainder of the topology.
 - While some operational differences are permissible, they should be minimized.

Dynamic Flooding – A Bit More

- One node (The Area Leader) is elected to compute the flooding topology for the dense subgraph.
 - Area leader election (generally) follows DR election semantics – with small differences
- This flooding topology is encoded into and distributed as part of the normal link state database.
 - Nodes within the dense topology would only flood on the flooding topology.
 - On links outside of the normal flooding topology, normal database synchronization mechanisms (i.e., OSPF database exchange, IS-IS CSNPs) would apply, but flooding would not
- Since the flooding topology is computed prior to topology changes, it does not factor into the convergence time and can be done when the topology is stable.
 - If a node has not received any flooding topology information when it receives new link state information, it should flood according to legacy flooding rules.

Simulation Results

- As predicted, a massive amount of improvement in flood reduction can be achieved with Dynamic Flooding optimizations
- As # of leaf, spine nodes increase, improvement approaches 95% reduction



Further Enhancements – Area Abstraction

draft-li-isis-area-abstraction

IS-IS Areas Are Transparent

- For traffic to transit level 1, some nodes and links must also be in level 2.
- In the limit, IS-IS areas do not aid scalability.
- Use case: Data centers/pods as level 1 areas.

Entire data center looks like a single node

Areas Should Be Atomic

- Abstract an area as a single node
- Use SR for transit connectivity

Multi-Level Abstraction



Area Hierarchy: IS-IS Multiple Levels 3-8

draft-li-isis-area-hierarchy

How do we hide the L1 topology from the rest of the network while preserving transitivity? Abstract the area itself into a single “node” representation

L3



L3

IS-IS PDU/Hello Header has reserved bits for 6 more levels of abstraction – allowing tremendous scalability (this isn't new – was already built in and we've seen this in PNNI)

Flooding radius bounded by areas – each area represents a multiplier in scalability

Summary of IS-IS “Native” Enhancements

- The point of IS-IS Dynamic Flooding, Area Abstraction, and Area Hierarchy work is to “upgrade” IGPs to the needs and network requirements of the 21st century
 - No desire to merge BGP and IGP capabilities
 - Goal is to keep them separate, attenuate BGP dump truck mentality
- Ultimately, dynamic flooding is just (dynamic) mesh groups, which have been around for 20 years
- Area hierarchy and abstraction allow for building end-to-end scale-out networks under a single IGP for topology discovery and dissemination
 - Topology independence, can be used anywhere



draft-ietf-rift...

RIFT (Routing in Fat Trees)

Overview of RIFT

- Background: DCs are largely Clos topologies, and thus admit to a potential optimization in how routing is done
 - Special topology, special protocol
- BGP used in DC with some hacks (RFC 7938)
 - AS numbering hacks, EBGP everywhere, ADD_PATHS, timer hacks, 4-byte ASNs, allow_as etc
 - Not ideal configuration-wise, but operators have instrumented their tools to support
- Link state information useful for enhancing routing, traffic steering, and resiliency
- Possible to merge the best of both Link State and Distance Vector approaches?

RIFT Summary

- RIFT aims to solve the DC routing problem by blending the best tenets of both link state and distance vector routing into a new hybrid style protocol
- It also aims to assist in configuration management by catering to autidiscovery and other autonomic networking needs
 - Scope limited to Clos topologies
- RIFT has gained enough industry interest that the IETF has chartered a working group to develop the technology
- Several talks have been dedicated to this topic, including at the past few NANOGs



LSVR (Link State Vector Routing)

IETF LSVR

Mostly from draft-ietf-lsvr-bgp-spf-02.txt

Overview of Link State Vector Routing

- Basic idea is to take advantage of BGP LS for information carriage and then run SPF computations on the resulting "LSDBs"
 - This is achieved by defining NLRI advertised within the BGP-LS/BGP-LS-SPF AFI/SAFI
- However, LSVR changes the Decision Process to allow for SPF computation
 - Particularly, Phase 1 and Phase 2 of BGP LS Decision process are replaced
 - Dijkstra algorithm run on LS info
- SPF Algorithm can be run in Strict or Normal Mode
- New Link, SPF NLRI proposed for BGP
- New Prefix and BGP-LS Sequence Number TLVs added to identify (and trigger SPF) on new LS info
- Protocol is modified to trigger decision process

LSVR Summary

- LSVR aims to solve the problem of IGP flooding scale limitations by leveraging BGP_LS as a transport for LS info and by modifying the BGP DP to understand this change and take advantage
- The upside is that a MSDC with existing BGP infrastructure and instrumentation can "upgrade" to this protocol and retain much of existing operations, at least theoretically
- Downside is that we are again blending two protocols and changing their respective behaviors in ways that may lead to unforeseen consequences
 - Always the case with new technology, so not a knock, just a fact



OpenFabric

draft-white-openfabric-06

Copyright © Arista 2018. All rights reserved.

ARISTA

Overview of OpenFabric

- The idea behind OpenFabric is to create a simple, autodiscoverable underlay routing capability
- Leverages IS-IS but with a lot of unnecessary stuff pulled out
 - External metrics (LSPs), TE extensions
- Adds some new capabilities in to streamline operations in L/S fabrics
 - Modified adjacency and optimized flooding mechanisms
- Designed only for Clos topologies, but can handle multiple stages
 - MUST NOT be mixed with standard IS-IS implementations in operational deployments!

OpenFabric Requirements

- OpenFabric is another proposal to leverage existing protocol technology in order to more easily build a datacenter without lots of configuration complexity
- Requirements are as follows:
 - Provide a full view of the topology from a single point in the network to simplify operations
 - Minimize configuration of each Intermediate System (IS) (also called a router or switch) in the network
 - Optimize the operation of IS-IS within a spine and leaf fabric to enable scaling



Observations and Comparisons

Subjective (mostly) and Open for Discussion!

Observations

- It is becoming obvious that there is interest in arriving at a protocol that **scales** extremely well, maintains **topology and link** behavior information, and **minimizes configuration** complexity
 - Policy control seems to be less desirable for underlay routing than expected
- **Scale-out design** principles are becoming more attractive
 - Disaggregated routing and cloud scale “follow me” adding momentum
- **Segment Routing** is showing up more and more as a patch/workaround or even elegant solution for some more difficult to address problems
 - Potentially an argument for the dreaded layer violation, but shows value overall of SR
- **Reducing protocols** in the network allows for more advanced protocol solutions without all the usual risk
- **Keep it Simple Sir**

Subjective Comparisons – Beer n’ Gear Debate Fodder

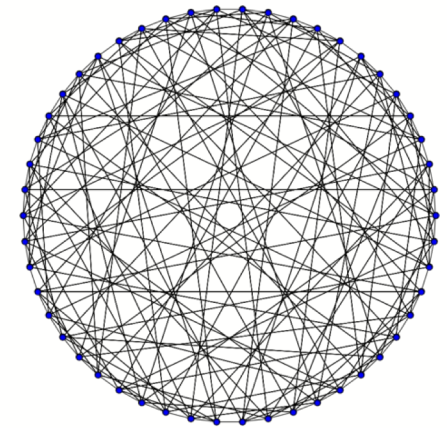
Protocol	SDO	Type	Scale	Config	Topology	Complexity	Notes
RIFT	IETF	Hybrid DV/LS	Unknown (high)	Autonomic	Clos	High	Early in dev, so some speculative Scale/Complexity. Borrows from IGP Native
OpenFabric	IETF	Modified LS	Medium	Light	Clos	Low	IS-IS as discovery protocol, borrows from IGP Native
LSVR	IETF	Hybrid DV/LS	High	High	Clos	Medium	Leverage BGP-LS, SPF, but hack decision process
IGP Dynamic	IETF	Extended LS	Very High	Light	Any	Medium	Complexity because of flooding topology algo and area abstraction

Other Approaches

- As expected, whenever there's a new technology gaining interest, there will be pile-ons
- SDN approaches to DCs, such as ONF's Trellis, aim to remove any notion of routing protocols from the equation
 - But the industry has been endowed with stable, well known protocols that are increasingly available freely – why disregard this exceptional bounty?
- Lots of folks pushing different ideas on how to tune IGPs to support Clos networks
 - le, draft-shen-isis-spine-leaf-ext-06
- Most approaches are subsets of the discussed approaches, and will likely see convergence into these approaches as we gear up for IETF 104 in Bangkok
- More will be revealed then! Stay Tuned

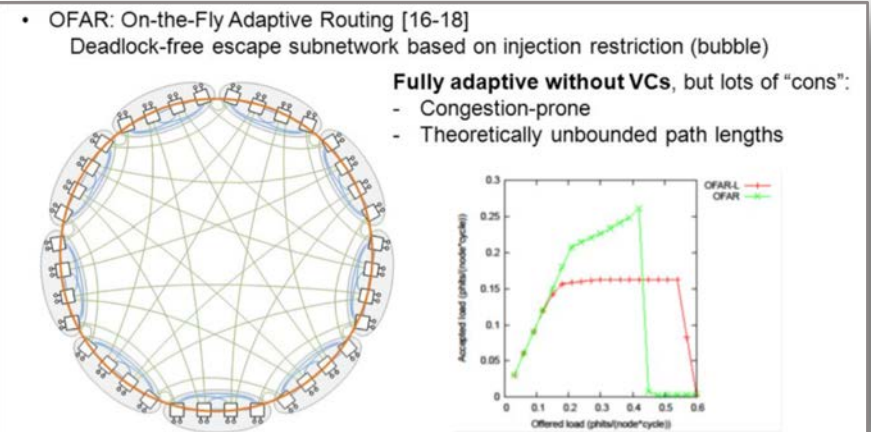
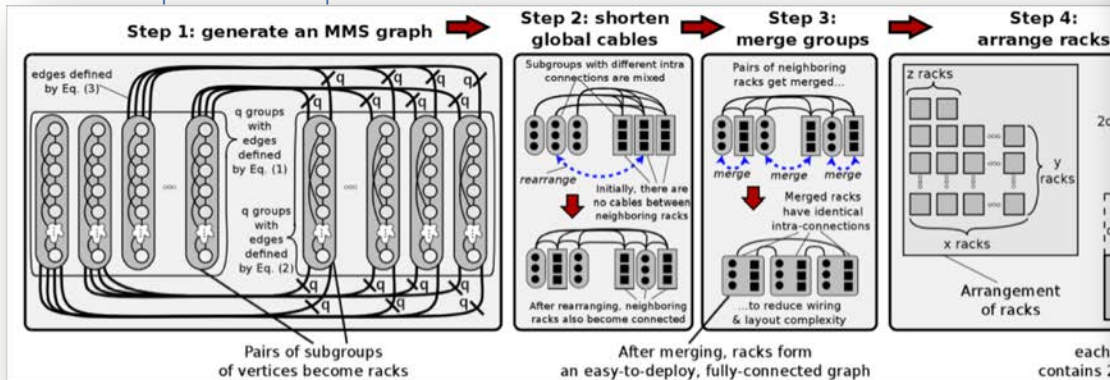
What of Non-Clos Topologies?

- Lots of novel topologies emerging in HPC – not long before they start showing up in DCs (especially low-latency)
- ECMP and full bisectional bandwidth vs full bandwidth under adversarial (non-uniform flows) are the trick
 - High radix routers change the game
- New routing algorithms possible on programmable silicon, SR allows for simplified VC paths
 - Significant gains in efficiency, if you dare!



Hoffman-Singleton graph is a 7-regular undirected graph with 50 vertices and 175 edges. It is the highest order Moore graph known to exist

*Moore Graphs bound optimal Slimflies





Conclusions

Copyright © Arista 2018. All rights reserved.

ARISTA

Concluding Remarks

- If the goal is to leverage as much of the benefits of LS IGPs as possible, then:
 - Fix IS-IS and OSPF flooding behavior
 - » IS-IS and OSPF both admit to fixing the flooding problem
 - Aim for additional scale via area abstraction and additional hierarchy
 - » IS-IS is best suited for this approach; OSPF hierarchy is fixed, but area abstraction may work
 - Open Fabric does some of this, IS-IS Dynamic/Area Abstraction takes it further
- If the goal is to leverage the best of LS IGPs and DV EGPs, then there are options here as well
 - RIFT, LSVR
- This author prefers a parsimonious approach either way
 - Limiting new technology development will likely save us from extended debate, low protocol uptake, and, very possibly, protocol correctness and stability problems
 - IGPs and EGPs are very old – and thus stable and well understood. Exploit!

Acknowledgements

I would like to acknowledge Tony Li for his significant contributions to the development of this technology and this presentation. I'd also like to acknowledge Keyur Patel, Tony Pryzgienda, Dave Cooper, and Alankar Sharma for their helpful discussions. Of course, none of this would matter without the contributions of the original authors and protocol dev teams who make all this work! To this end, I'd like to acknowledge Peter Psenak, John Drake, Les Ginsberg, and Eric Rosen for their discussions and ongoing contributions, and to Russ White and Shawn Zandi for their work on OpenFabric.



Thank You!

Questions? Comments?
cmartin@arista.com

Copyright © Arista 2018. All rights reserved.

ARISTA