IGPs - the new hotness

CHI-NOG 09 :: 23-may, 2019

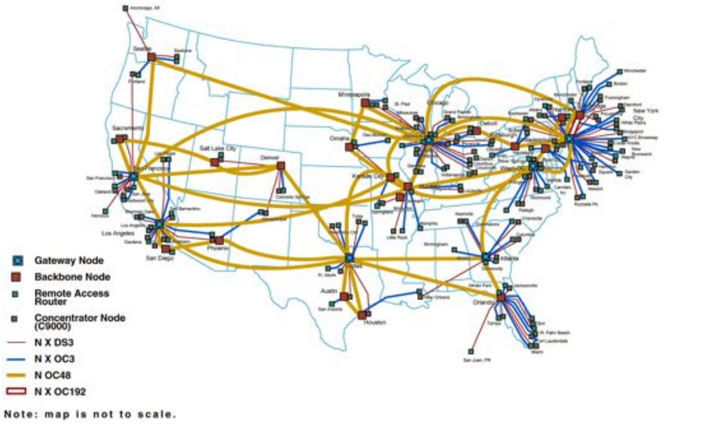
steve ulrich (sulrich@arista.com)

agenda

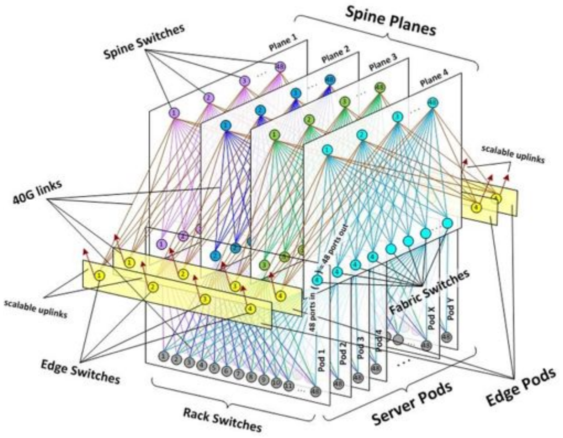
- background
- recent DC protocol activities
  - RIFT
  - OpenFabric
  - LSVR
- a generalized approach
  - draft-ietf-lsr-dynamic-flooding
  - draft-li-lsr-isis-area-abstraction
  - draft-li-lsr-isis-hierarchy
- conclusion

background

- platform forwarding densities and changes in network design are changing the scaling requirements of IP and MPLS networks
  - flat, partial mesh topologies >> hierarchical clos networks



c/o "An Atlas of Cyberspace"
https://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/atlas/more_isp_maps.html

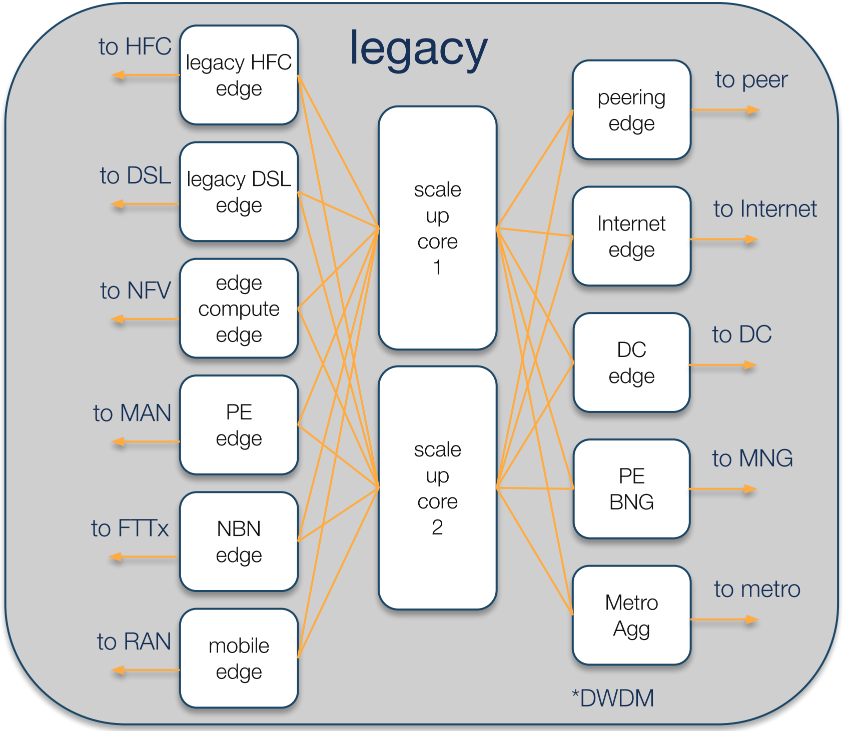


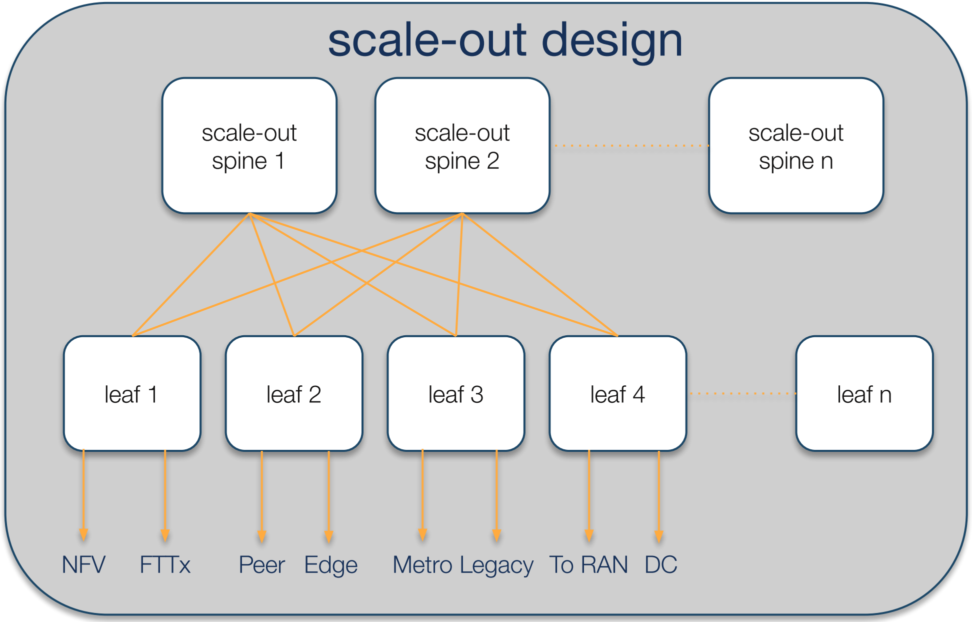
c/o Facebook Engineering
"http://code.facebook.com"

- IGP scaling limitations have led to BGP adoption as de facto protocol for DC
  - using BGP in this application overloads the semantics of the protocol
  - loss of topology detail reduces value of IGP-based forwarding mechanisms (i.e. LFA)

scale up vs. scale-out

scale-out isn’t a significant change from existing carrier topology or traffic flows. however, it is more efficient.





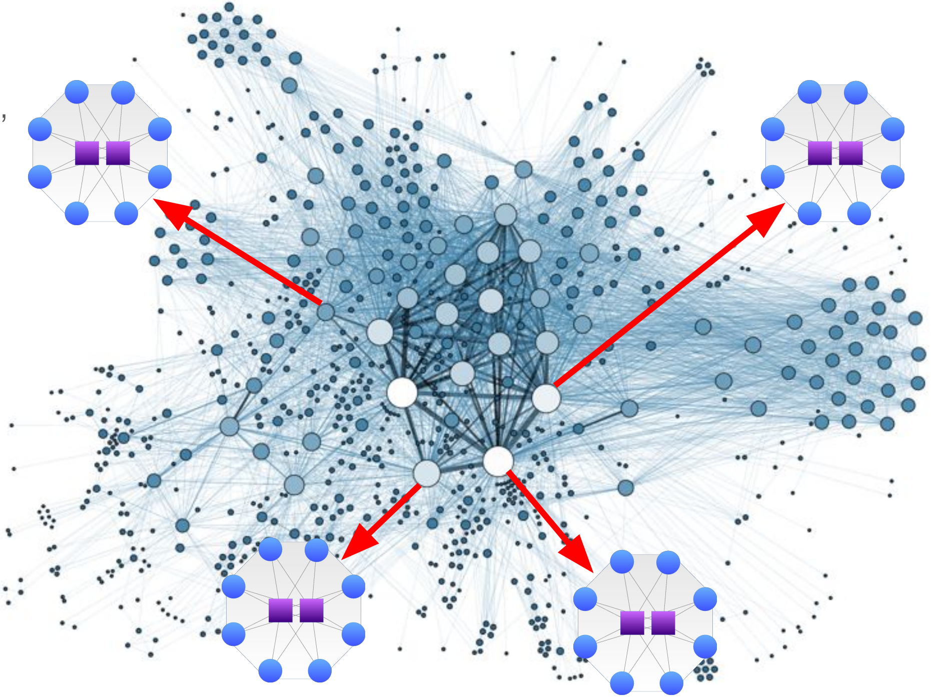
- ❌ site capacity limited by scale of core routing platforms
- ❌ difficult to take devices out of service for maintenance
- ❌ single purpose edge routers Increases CAPEX and core intf count

- ✔ elastic site capacity – add spine or leaf nodes
- ✔ simplified operation with higher availability
- ✔ scale-out reduces CAPEX, enables true pay as you go

challenges with the current approach

- IGP scaling issues have (largely) led to widespread use of BGP in DCs
- BGP adoption in MSDC is widespread
  - BGP app dev for policy control via tooling (ie, BIRD, Quagga, GoBGP)
  - simple configuration if automated
  - known scale, symmetric topology; yields simplified deployment
  - ECMP L/S design reduces convergence scope to clos stage width
- BGP can’t really be used as an IGP in an arbitrary topology without significant amount of configuration
  - difficult to automate config w/irregular topologies
  - single router per AS = RIPv4

we need a solution to networks like these, where the circular nodes may themselves be comprised of dense graphs



trends in large scale systems

disaggregated, deconstructed and distributed

- evolving towards routing “supernodes” built of leaf / spine elements
- need to make the guts of a supernode have minimal impact on overall network scale
- enables complete fungibility and best of breed “line card” selection

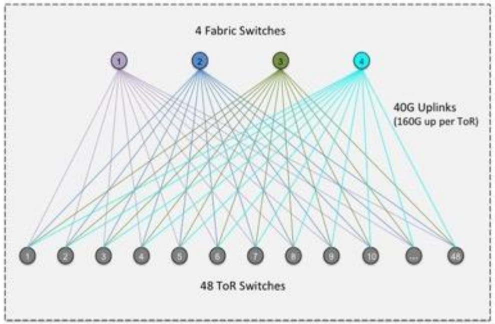
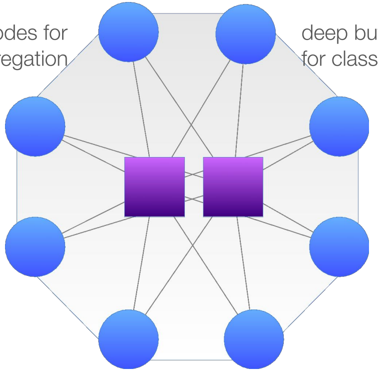


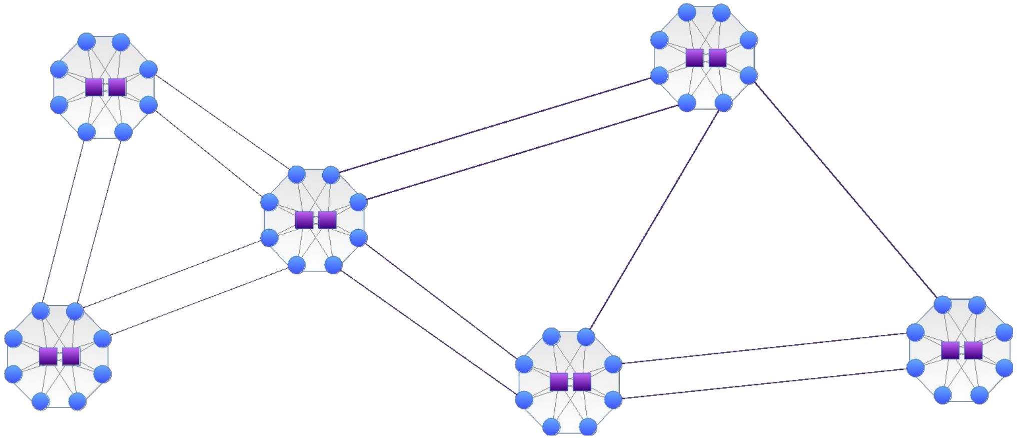
Figure 2: The internal pod topology of the Altoona architecture



thin-FIB / buffer nodes for L2 switching, aggregation

deep buffer high-scale FIB nodes for classic Edge/Peering

high-touch, programmable forwarding nodes for edge/service applications



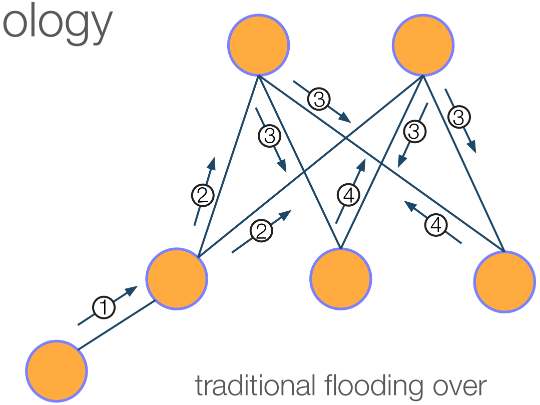
animated

IGP flooding example

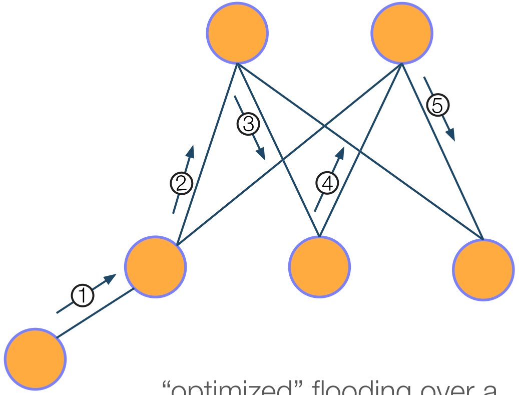
- IGP flooding is opportunistic and complete

  flood everywhere while maintaining transmission lists to prevent endless reflooding, w/split horizon

- in dense, symmetric graphs, the amount of information flooded overwhelms the control plane at scale, with no solution (to date) other than avoidance
- objective is to reduce flooding to a minimal (not necessarily optimal) flooding topology



traditional flooding over full graph~ O(n^2)



"optimized" flooding over a minimum vertex cover ~ O(n)

recent (DC) protocol activities

- RIFT
- OpenFabric
- LSVR

Routing In Fat Trees (RIFT)

draft-ietf-rift-rift

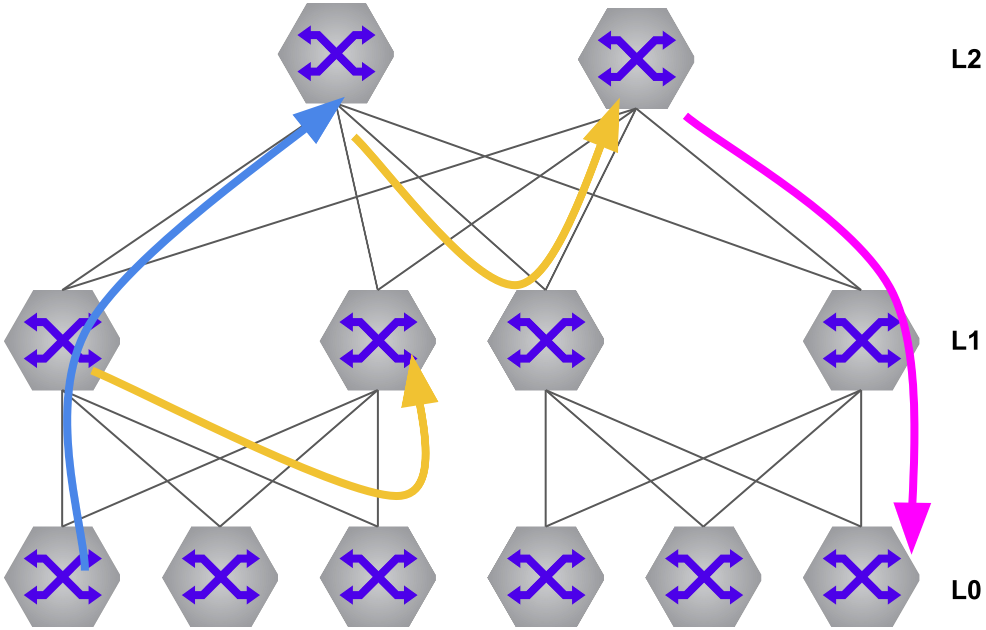
- DCs are largely symmetric topologies
- with a predictable topology lots of optimization is possible
  - special topology, special protocol
- link state information useful for enhancing routing, traffic steering and resiliency
- RIFT attempts to merge the best of link state and distance vector approaches
- IETF working-group

2 implementations available

- juniper trial implementation
- python implementation (github)

RIFT operational overview

- **link-state flood up**
  - full topology and all prefixes at top level only
- **distance vector down**
  - 0/0 is sufficient to send traffic up.
  - more specifics
    - disaggregated in case of failure
    - PGP for traffic steering
- **bounce**
  - flooding reduction
  - automatic dis-aggregation



OpenFabric overview

draft-white-openfabric-XX >> draft-white-distoptflood-00

- OpenFabric creates a simple, auto-discoverable underlay routing capability
- leverages IS-IS removes unnecessary elements
  - external metrics (LSPs), TE extensions
- adds some new capabilities in to streamline operations in L/S fabrics
  - modified adjacency and optimized flooding mechanisms
- designed specifically for clos topologies
  - but it can handle multiple stages
  - **MUST NOT** be mixed with standard IS-IS implementations in operational deployments!

Free Range Routing (FRR) implementation available

Link State Vector Routing (LSVR) overview

- applicability doc
- BGP SPF extensions

- take advantage of BGP-LS for topology distribution and then run SPF computations on the resulting "LSDB"
    - achieved by defining NLRI advertised within the BGP-LS/BGP-LS-SPF AFI/SAFI
- LSVR changes the BGP-LS decision process to allow for SPF computation
    - phase 1 and phase 2 of BGP-LS decision process are replaced
    - dijkstra algorithm run on L-S info
- SPF algorithm can be run in strict or normal mode

- defines new NLRIs for BGP
    - Link
    - SPF NLRI
- new TLVs are defined to identify new link-state information and trigger SPF calc
    - prefix TLV
    - BGP-LS sequence umber TLVs
- protocol is modified to trigger decision process

LSVR Summary

- LSVR addresses IGP flooding scale limitations by using BGP-LS as a transport for L-S info and modifying the BGP decision process
- an MSDC with existing BGP infrastructure and instrumentation can potentially "upgrade" to this protocol and retain much of existing operations
- two protocols are getting blended and behaviors are getting changed in ways that may lead to unforeseen consequences

a more generalized approach …

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

link state protocols - what do ya get?

- info re: topology state and link characteristics in the network is easily conveyed in the IGP
  - these are used for critical forwarding plane operations
- next generation multicast (BIER) and traffic engineering (Segment Routing and RSVP) benefit from a L-S IGP
  - in the absence of a controller and detailed topology discovery, it's the only way to do Segment Routing, RSVP-TE, and BIER
  - TI-LFA is critical for ensuring resilience without RSVP-TE FRR (which, btw, requires an L-S IGP)
- the need 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)
  - reference challenges re: inter-area link/node protection, inter-AS TE, etc.

link state protocols - dynamic flooding

- in a dense topology, the flooding algorithm that is the heart of conventional link state routing protocols has a large amount of redundant messaging.
    - this is amplified by scale.
- the protocol (when judiciously deployed) survives this combination, the redundant messaging is unnecessary overhead and delays convergence.
- the problem is to provide routing in dense, scalable topologies with rapid convergence.
- we need a flooding topology that is a subset of the forwarding topology

link state protocols - dynamic flooding requirements

draft-ietf-lsr-dynamic-flooding

1. provide a dynamic routing solution
   - reachability must be restored after any topology change
2. provide a significant improvement in convergence
3. must address a variety of dense topologies.
   - solely addressing a complete bipartite topology is insufficient
   - multi-stage clos topologies (and variations) must also be addressed
   - addressing complete graphs is a good demonstration of generality
4. there must be no single point of failure
   - the loss of any link or node should not unduly hinder convergence
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.

link state routing - dynamic flooding (continued)

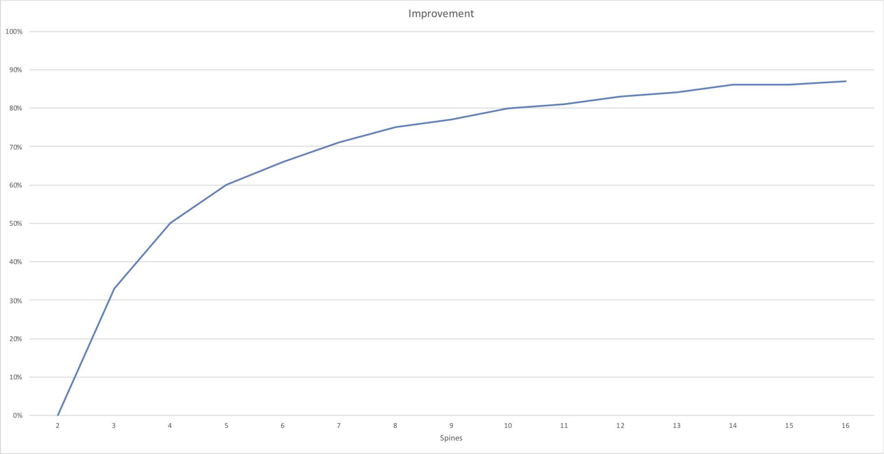
draft-ietf-lsr-dynamic-flooding

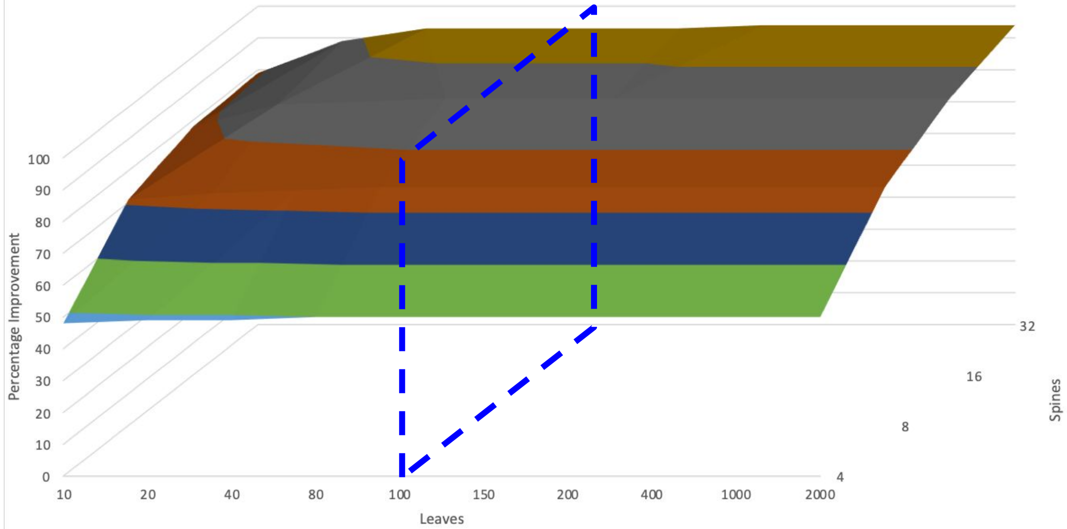
- one node (area leader) elected to compute the flooding topology for the dense subgraph
  - area leader election largely follows DR election semantics – with small differences
- flooding topology is encoded into and distributed as part of the normal link state database
  - nodes within the dense topology 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
- 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

draft-ietf-lsr-dynamic-flooding

- a massive improvement in flood reduction can be achieved with dynamic flooding optimizations in dense graphs
- as # of leaf, spine nodes increase, improvement approaches 95% reduction in flooding overhead
- current WAN deployments also see notable flooding reduction improvements, though not as significant
  - study of various backbones
  - mean flooding reduction of ~30%
  - highest: 90% flooding reduction
  - lowest: none





area abstraction

draft-li-lsr-isis-area-abstraction

IS-IS areas are transparent

- for traffic to transit level 1, some nodes and links must also be in level 2.
- IS-IS areas do not aid scalability.
- use case: data centers/pods as level 1 areas.

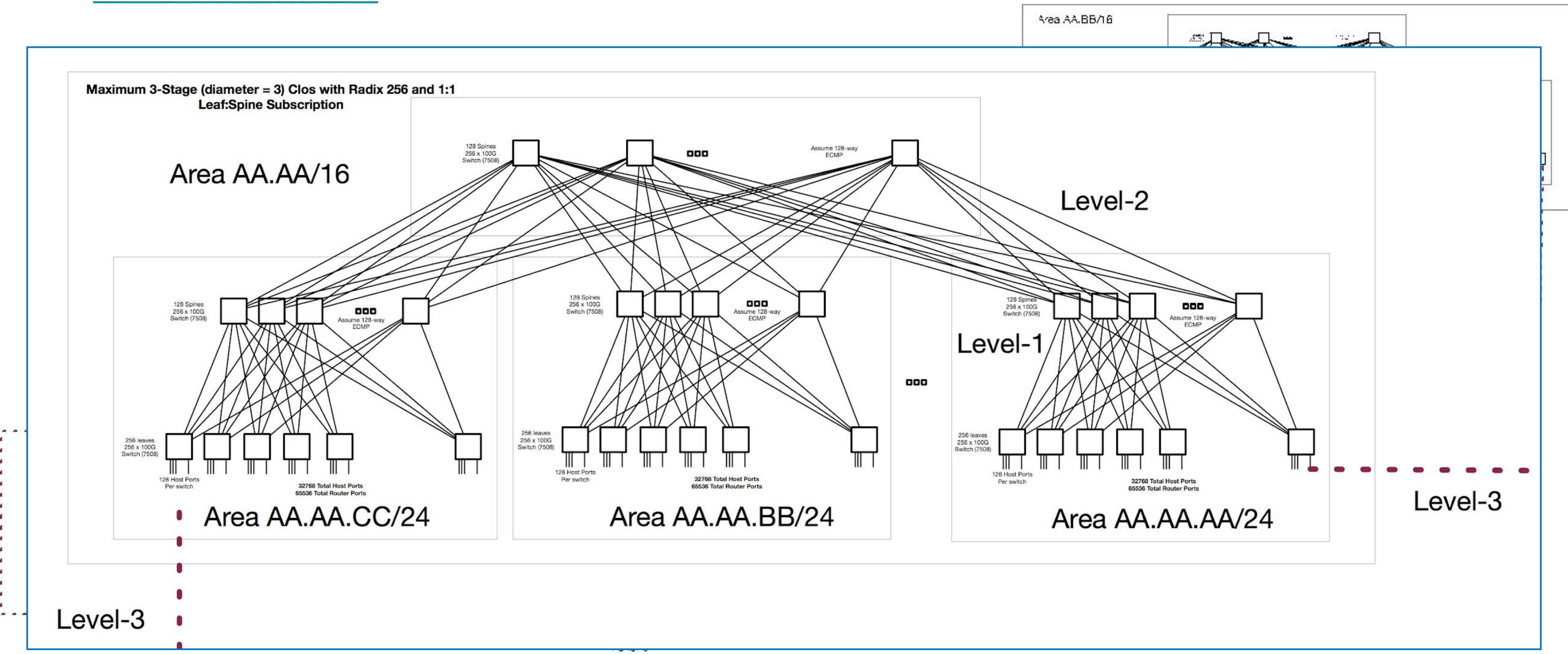
IS-IS areas should be atomic

- abstract an area as a single node
- use segment routing for transit connectivity

make an entire data center look like a single node animated

scale-out network design

draft-li-lsr-isis-area-abstraction


Area AA.BB/16
Maximum 3-Stage (diameter = 3) Clos with Radix 256 and 1:1 Leaf:Spine Subscription
Area AA.AA/16
128 Spines 256 x 100G Switch (7508)
Assume 128-way ECMP
Level-2
128 Spines 256 x 100G Switch (7508)
Assume 128-way ECMP
128 Spines 256 x 100G Switch (7508)
Assume 128-way ECMP
128 Spines 256 x 100G Switch (7508)
Assume 128-way ECMP
Level-1
256 leaves 256 x 100G Switch (7508)
128 Host Ports Per switch
32768 Total Host Ports 65536 Total Router Ports
Area AA.AA.CC/24
256 leaves 256 x 100G Switch (7508)
128 Host Ports Per switch
32768 Total Host Ports 65536 Total Router Ports
Area AA.AA.BB/24
256 leaves 256 x 100G Switch (7508)
128 Host Ports Per switch
32768 Total Host Ports 65536 Total Router Ports
Area AA.AA.AA/24
Level-3
Level-3

area hierarchy: IS-IS Multiple Levels 3-8

draft-li-lsr-isis-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

- IS-IS PDU/Hello Header has reserved bits for 6 more levels of abstraction
  - there's potential for significant growth in the scalability domain
  - 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

IS-IS enhancements

- 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
  - keep them separate
- 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

conclusion

comparison (subjective)

protocol	SDO	type	scale	config overhead	topo	complexity	notes
RIFT	IETF	hybrid DV/LS	unknown (high)	autonomic	clos	high	- under development - scale unknown - complexity unknown - borrows from IGP
OpenFabric	IETF	IS-IS derivative	medium	light	clos	low	- IS-IS as discovery - borrows from IGP
LSVR	IETF	hybrid DV/LS	high	high	clos	medium	- leverage BGP-LS, SPF - modify decision process
IGP dynamic	IETF	extended LS	high	light	any	medium	complexity from flooding topology algo and area abstraction

conclusions

- there is interest in arriving at a protocol that scales extremely well, maintains (and distributes) 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
    - foster a whole range of new deployment models and network architectures
- reducing protocols in the network allows for more advanced protocol solutions while limiting risk

thank you!