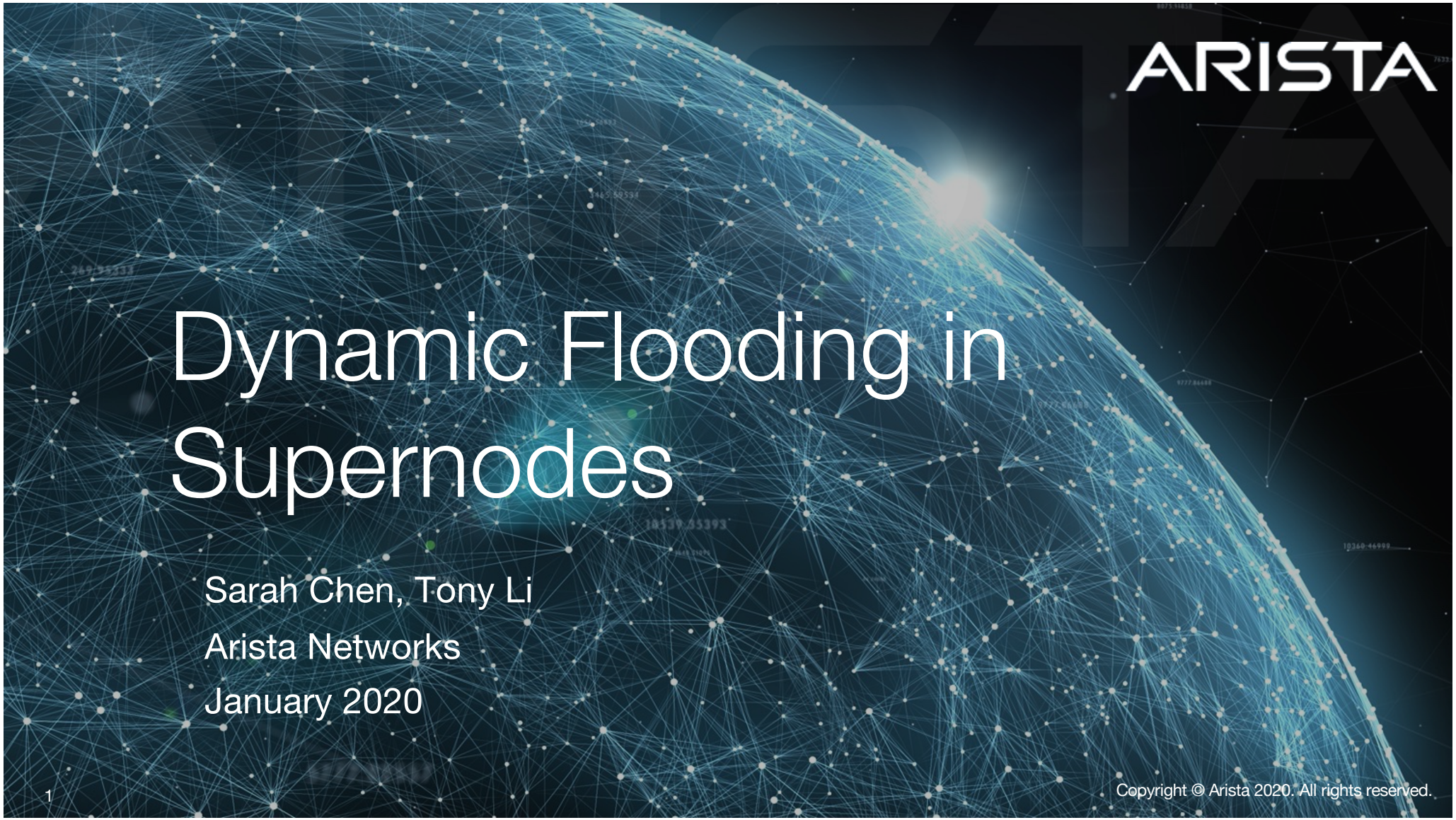


Dynamic Flooding in Supernodes

Sarah Chen, Tony Li
Arista Networks
January 2020

Introduction

- Supernode Architecture
  - Supernodes are alternatives to core routers.
  - Link state IGP needed for TE, SR, BIER.

- Improve link state IGP flooding in dense topologies.
  - Dynamic flooding.

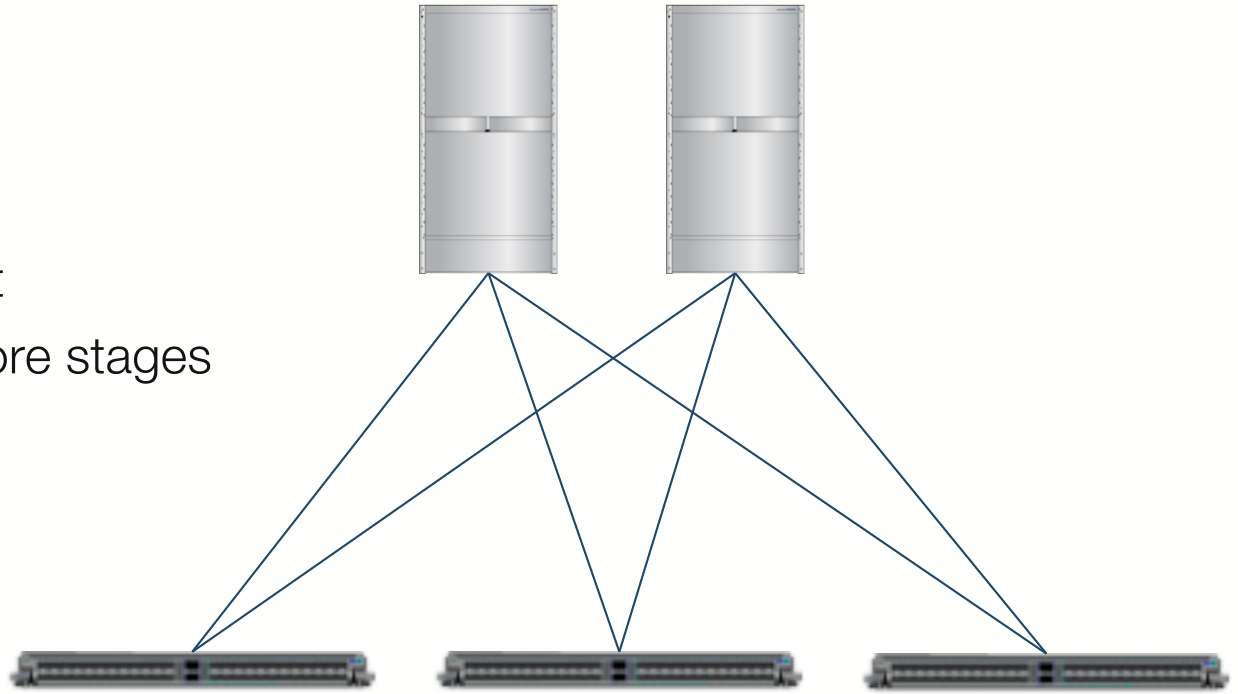
What's in a core router, anyway?

- Fabric – provides bandwidth
  - What if it was Ethernet instead?
- CPU – Control and management
  - What if control and management was distributed?
- Line cards – Interfaces plus packet forwarding
  - What if they were small, fixed form factor routers?

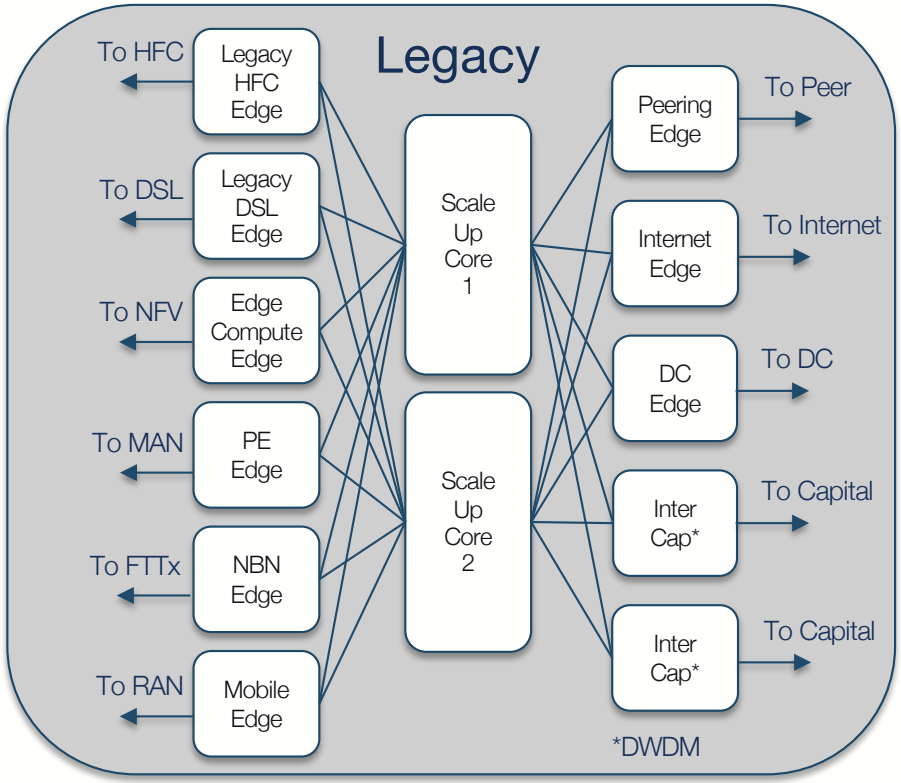
Building supernodes

- Fabric → Dual spines
- Line cards → Small routers

- Leaf-spine topology to start
- Scale with more spines, more stages

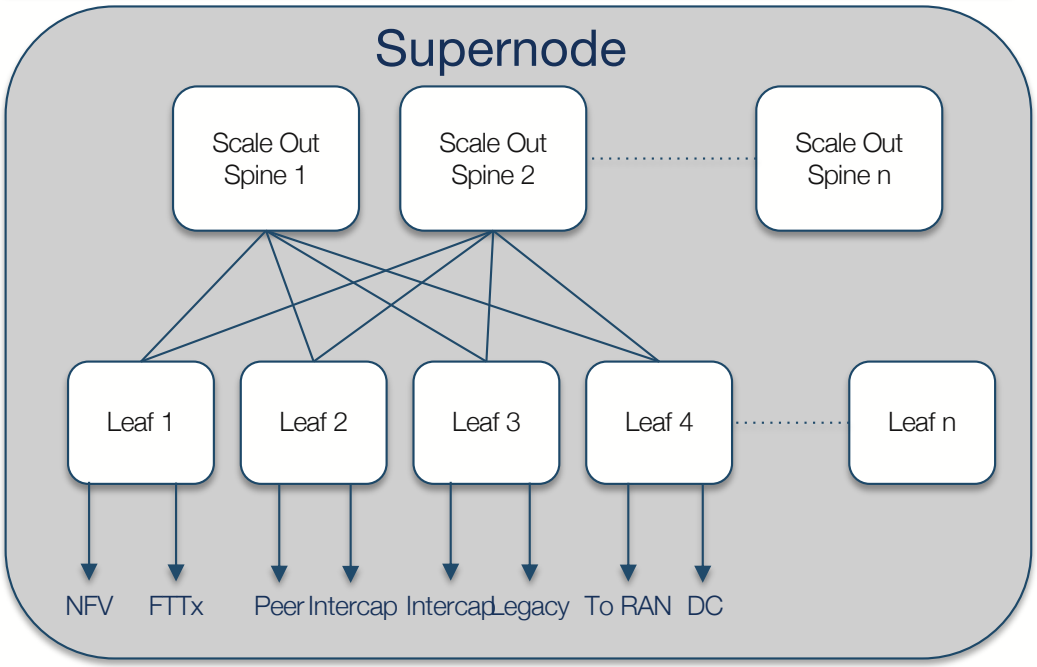


Scale Up Vs Scale Out



Site Capacity Limited By Max Scale of one Core Router
Difficult to take Core/Edge Routers Out of Service for Maintenance
Large 'blast radius' for Core Routers, increased Software Complexity

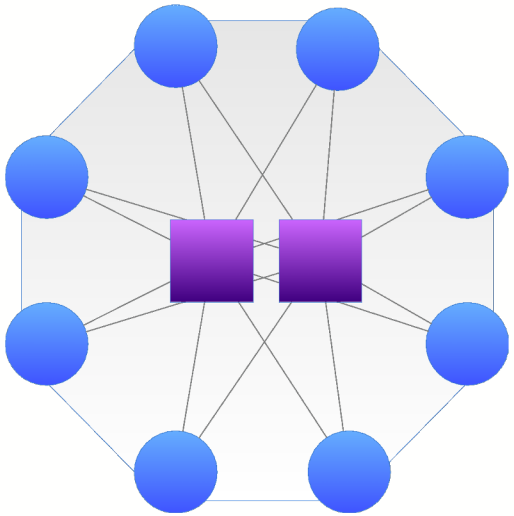
Scale Out Isn't A Radical Change From Existing Carrier Topology Or Traffic Flows. It Is More Efficient.



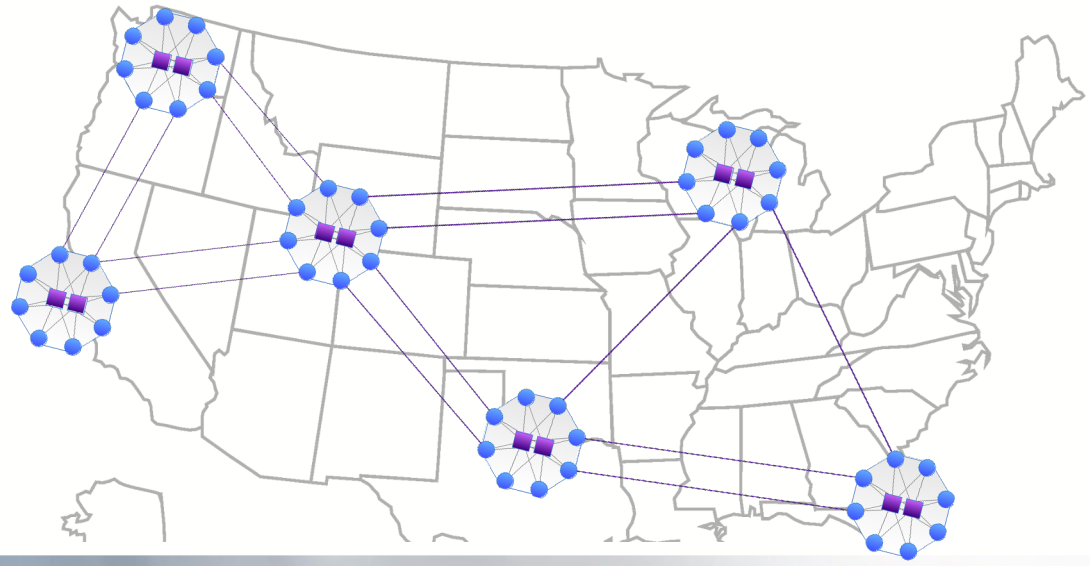
Elastic Site Capacity – Add More Spines or Leaf Nodes
Ease of Service and Upgrade – Incremental Per Node
Scale Out Reduces CAPEX, Converges Edges, Provides Optionality

ARISTA

Supernode Architecture in the Large

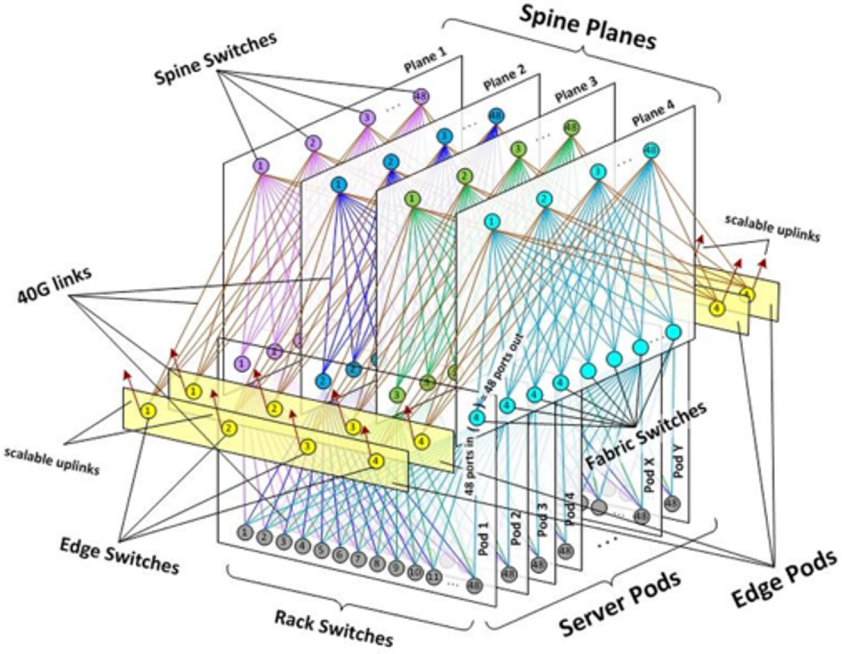


- Routing “Supernodes” built of leaf/spine topologies
- Key is to make scaling “inside” the supernode to have minimal impact on total network scaling
- Benefit extends beyond avoiding multichassis system design – complete asset fungibility and best of breed “linecard” selection with rolling upgradeability



Supernode To-do List

- Scalable routing within the supernode
  - Historically, LS routing has been done with BGP. Tractable only by automation on a very strict topology.
  - Link state flooding on dense topologies needs improvement.
- Control plane abstraction
  - One node outside of the boundary
  - Better IGP abstraction
- Architectural Scalability
  - Additional IS-IS levels – Two are not enough



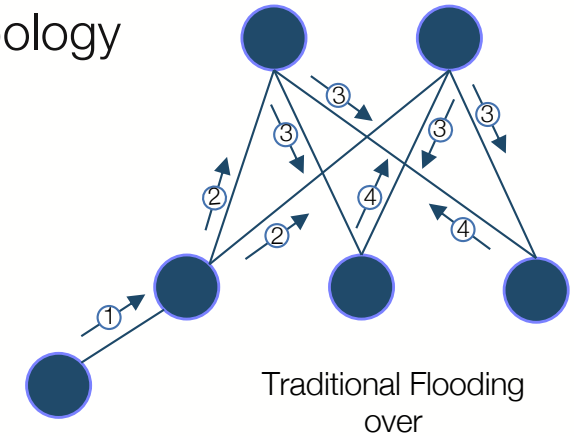
Why Link State?

- 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) benefits 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

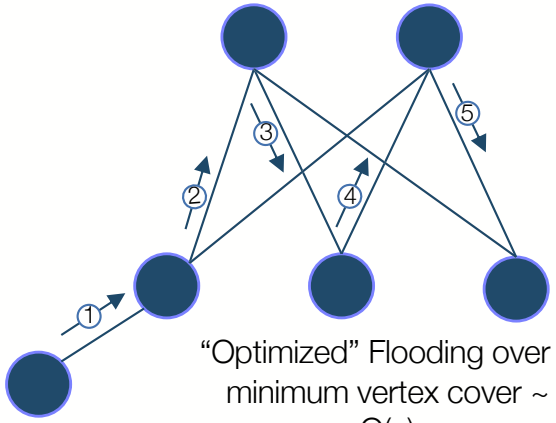
IGP Flooding Example

*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 is to reduce flooding to a minimal (not nec. optimal) flooding topology



Traditional Flooding over full graph~ O(n^2)



"Optimized" Flooding over a minimum vertex cover ~ O(n)

draft-ietf-lsr-dynamic-flooding

Dynamic Flooding

ARISTA

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 should address a variety of dense topologies.
  - Just addressing a complete bipartite topology such as K5,8 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 Centralized Approach

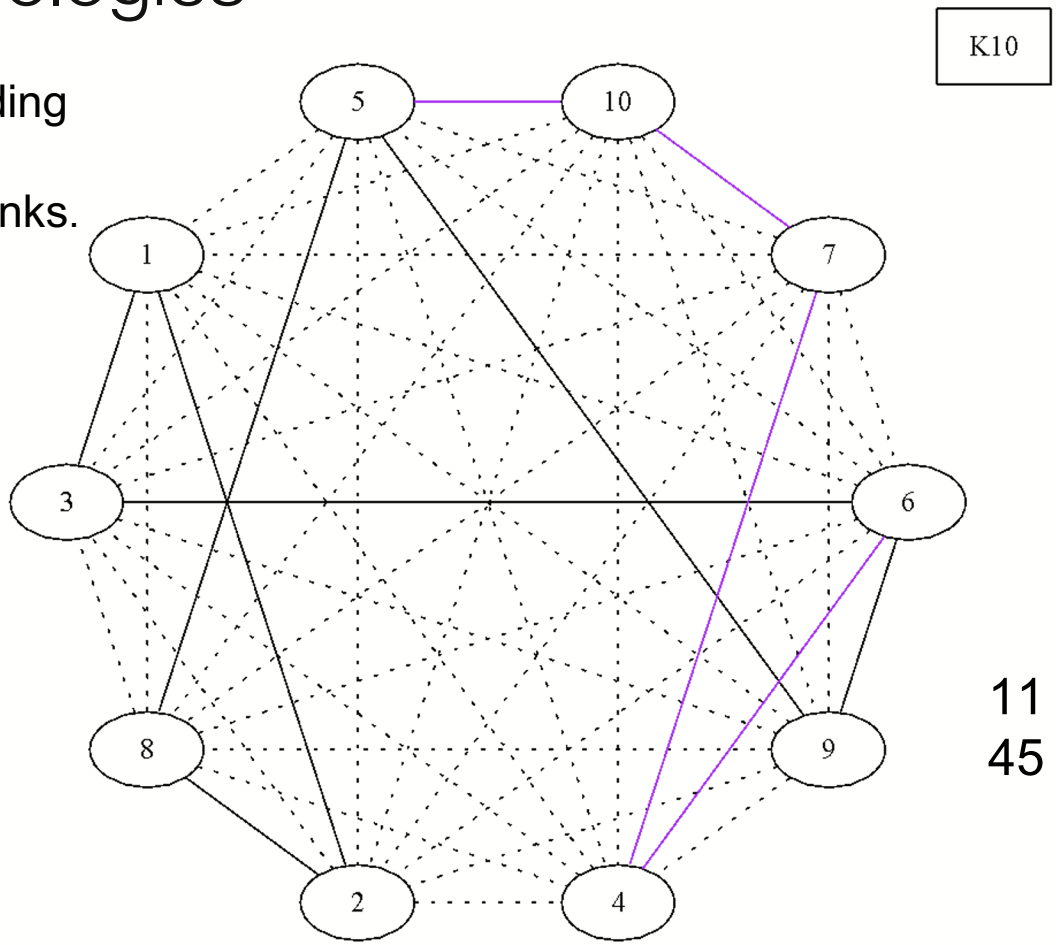
- One node (The Area Leader) is elected to compute the flooding topology for the dense subgraph.
  - Area leader election (generally) follows DR election semantics
- This flooding topology is encoded into and distributed as part of the normal link state database.
  - Nodes within the dense topology would only flood on the flooding topology.
  - Normal database synchronization mechanisms (i.e., OSPF database exchange, IS-IS CSNPs) still apply on all links.
- The flooding topology is updated by the Area Leader upon topology changes.
- Nodes can request temporary flooding from neighbors.
  - Add links to the flooding topology temporarily in some topological events.

Protocol Extension

- New TLVs for IS-IS and OSPF:
  - Area Leader sub-TLV
    - » indicate a system's preference for becoming Area Leader.
  - Dynamic Flooding sub-TLV
    - » indicate that it supports Dynamic Flooding and the algorithms that it supports for distributed mode, if any.
  - Area Node IDs TLV
    - » carry the list of system IDs that compromise the flooding topology for the area.
  - Flooding Path TLV
    - » carry a path which is part of the flooding topology
  - Flooding Request TLV
    - » request flooding from the adjacent node

Flooding Topologies

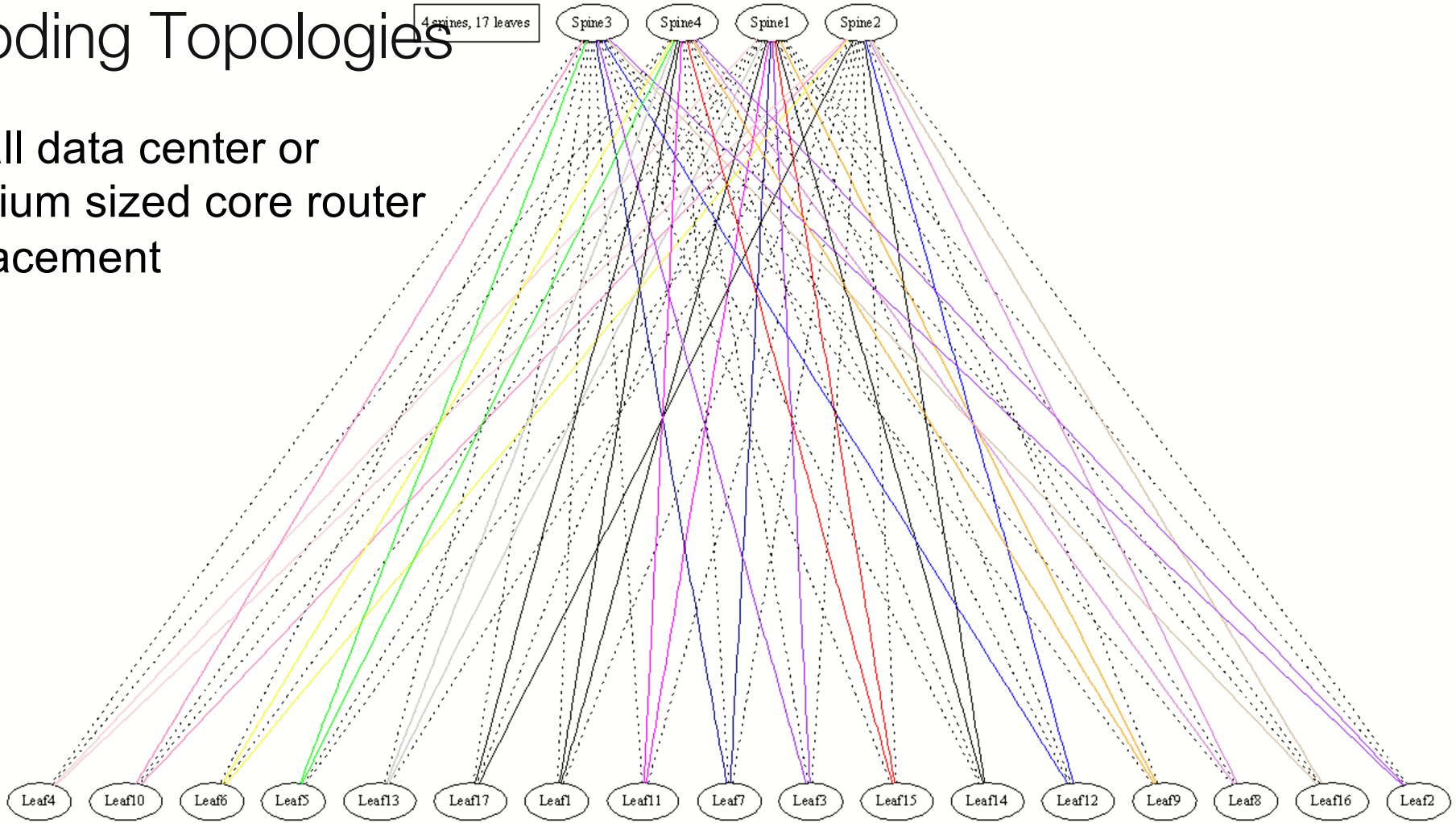
Solid lines are the flooding topology. Dashed lines are non-flooding data links.



11 flooding links
45 data links

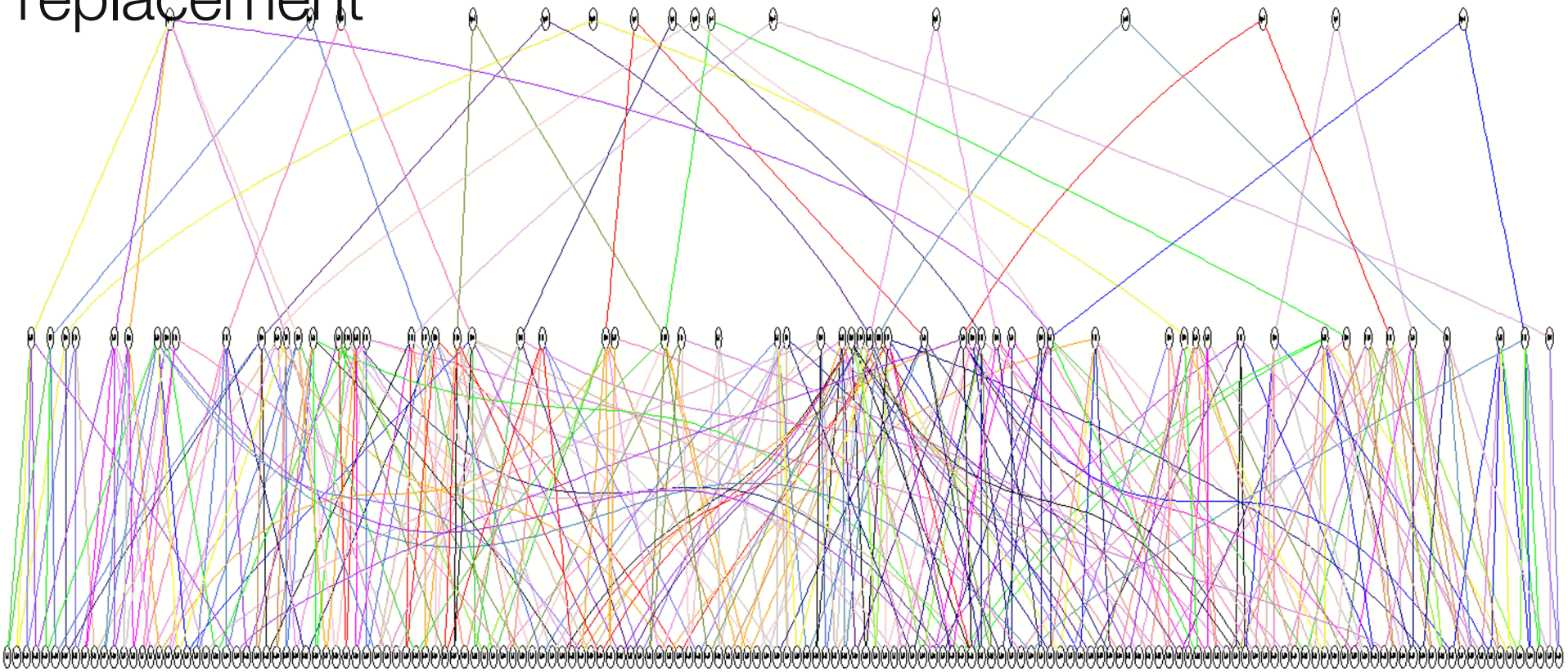
Flooding Topologies

Small data center or
Medium sized core router
replacement

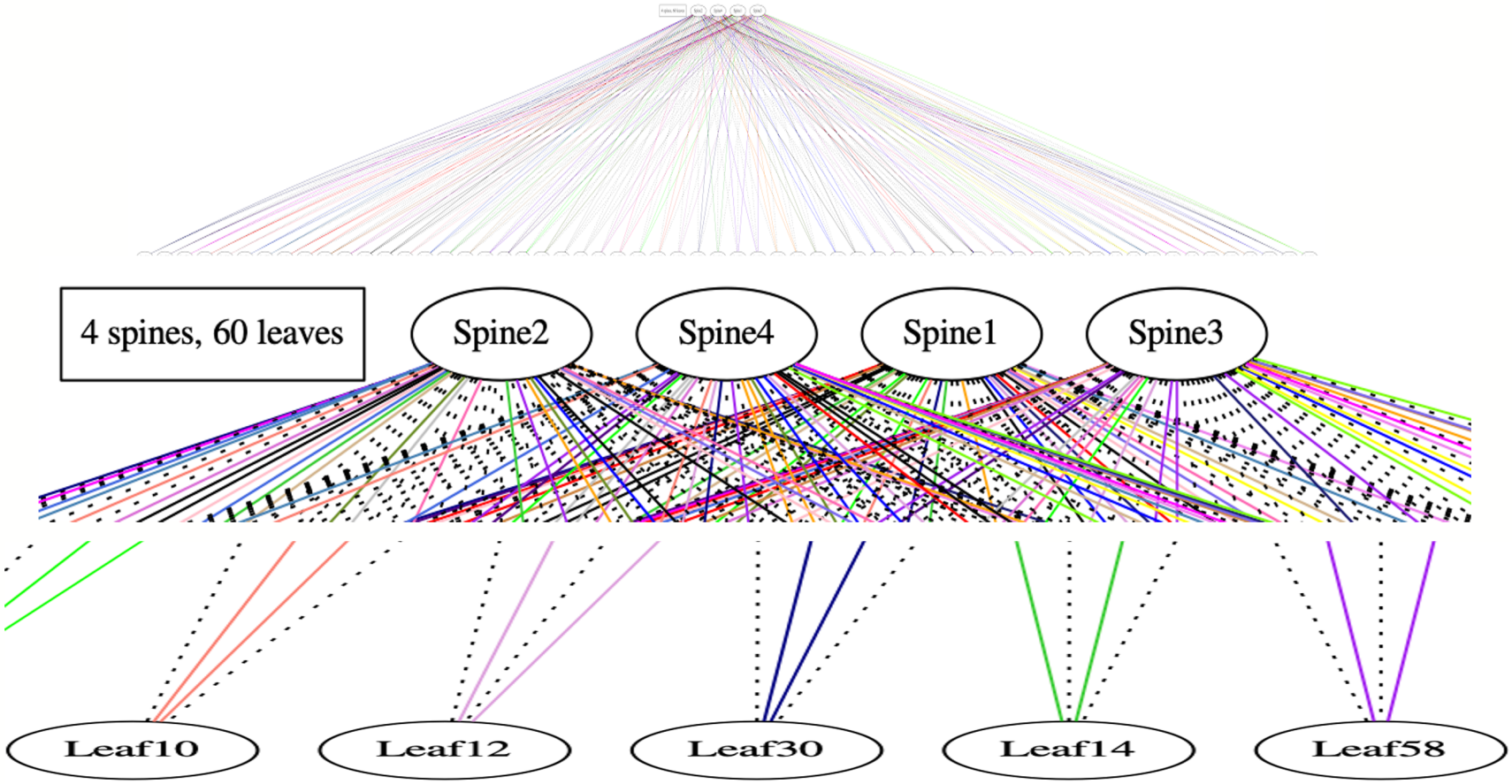


ARISTA

Flooding Topologies – $K_{16,64,128}$ – Large core router replacement



Flooding Topologies – $K_{4,60}$



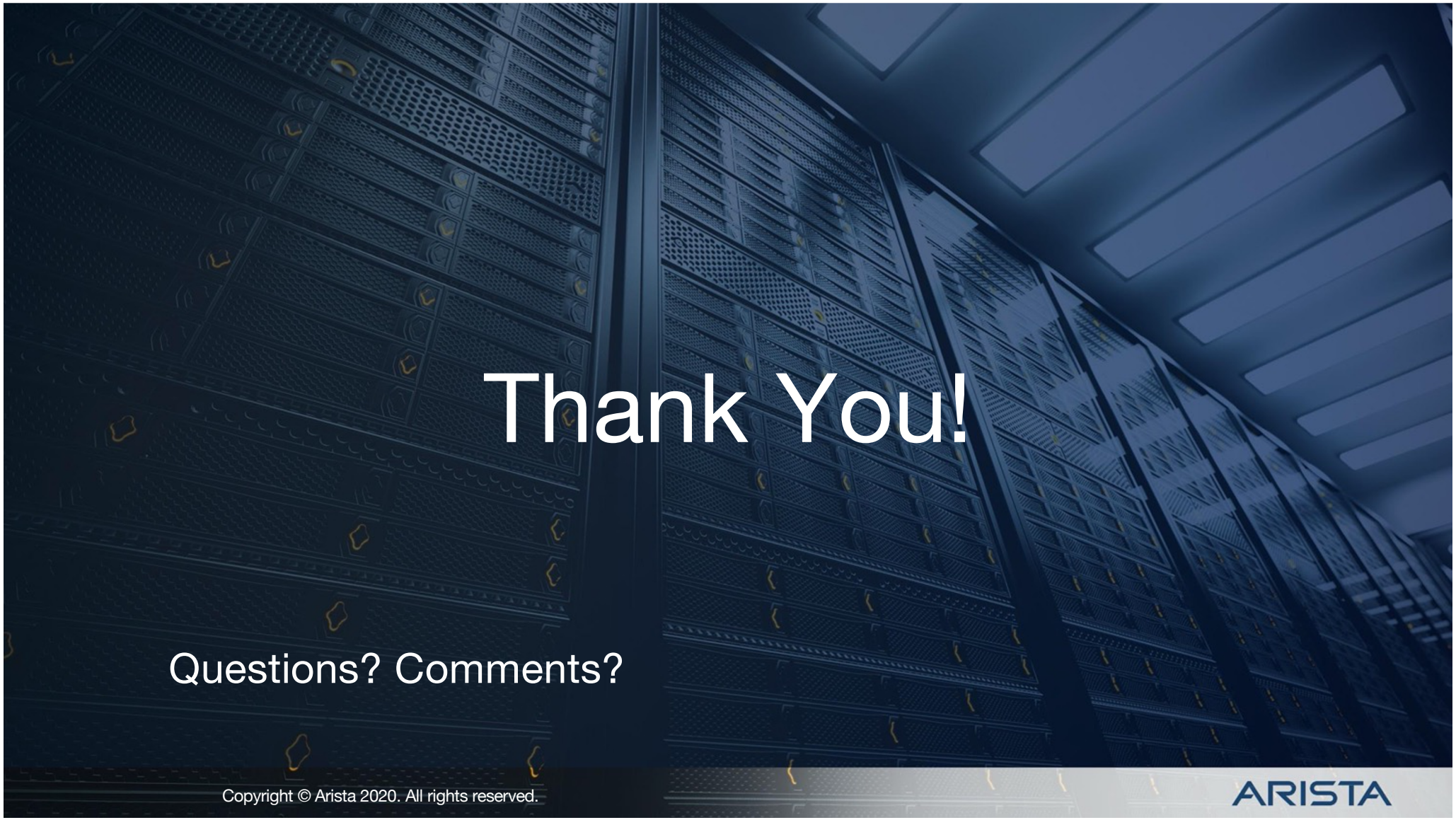
Results

- Boot and stabilize $K_{4,60}$
- Count LSPs & IIHs

	Original	Dynamic Flooding
Spine	110,000	41,900 (-62%)
Leaf	15,000	3,350 (-77%)

Summary

- Dynamic Flooding dramatically reduces the number of links used for flooding in dense topologies.
- Decreased flooding allows the control plane to scale.
- Using large topologies, we can build supernodes that are economical replacements for legacy core routers.
- Supernodes:
  - Multi-vendor – Competitive pricing
  - Mix-and-match chassis – Buy what you need, when you need it.
  - Elastic capacity – Grow as you need, in small increments if necessary.
  - Scalable – Grow as much as you need. Technology independent.
  - Upgradeable – Rolling single chassis upgrades are less disruptive.
  - Safer – Failures are localized to the chassis.



ARISTA