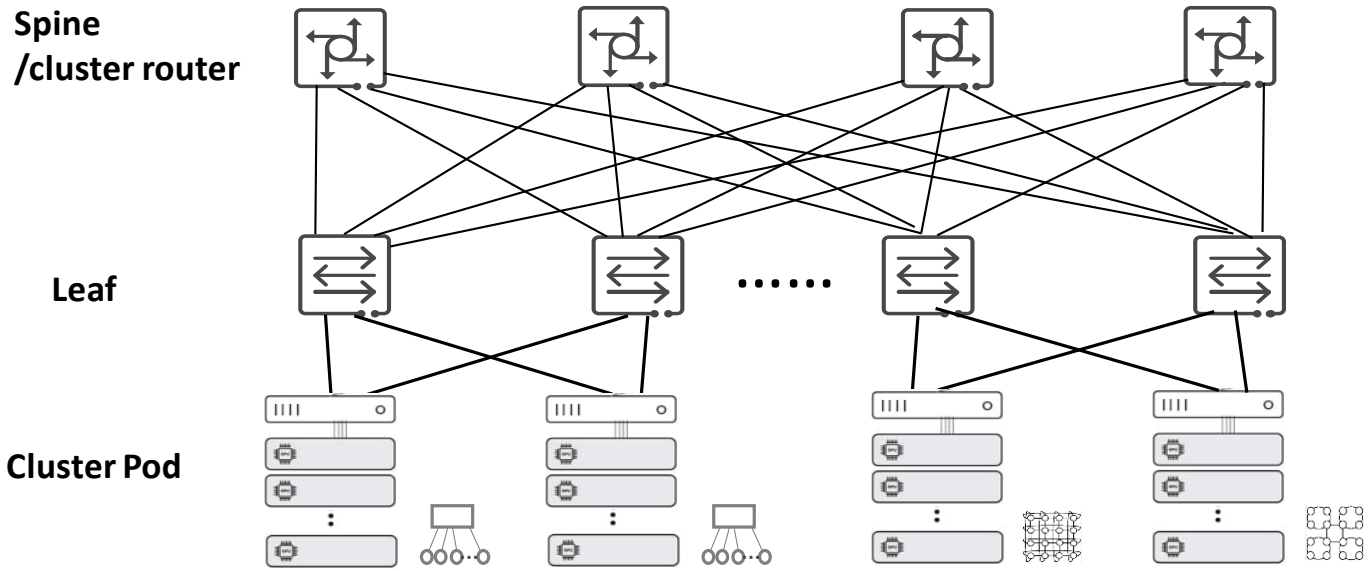
The Challenges and Requirements for Routing in Computing Cluster network

draft-li-rtgwg-computing-network-routing

IETF121 RTGWG, Dublin

Fengkai Li & Yizhou Li

Background



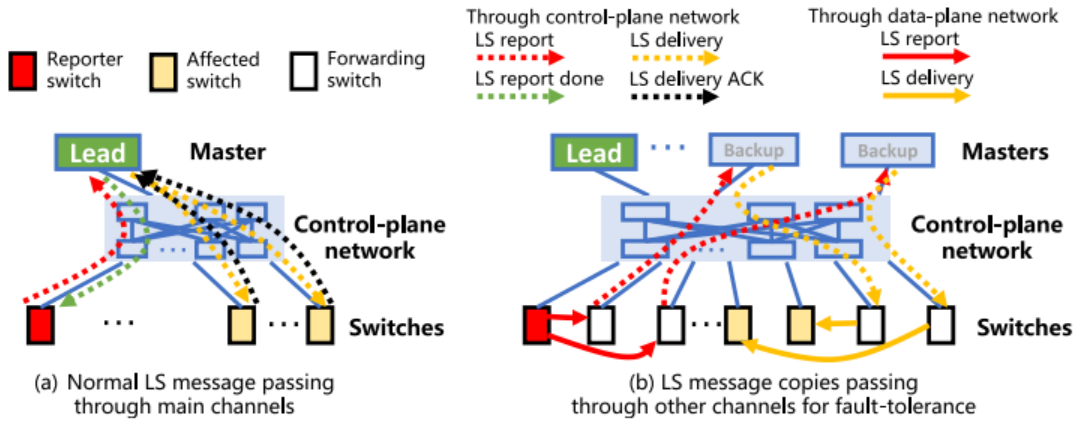
- **Some well-known characteristics**
  - Large scale, multiple path available
  - Prior knowledge of the topology, paths and traffic patterns
  - Provide high bandwidth and low latency for large data volume
  - High modularity, the system is composed of clusters with known structures
- **Key considerations on routing in computing cluster network**
  - Scalability, availability
  - Failure handling and route/path convergence
  - Complexity of network configuration and maintenance

Routings in computing cluster network

- **Distributed routing**
  - Different flavors using BGP, e.g. [BGPinDC], [HPN]…
  - Pros: Good scalability, flexibility of policy control
  - Cons: Heavy configurations and slow routing convergence (per-node calculation and propagation)
- **Centralized routing**
  - SDN-based routing infrastructure, e.g. [ORION], Borg [BORG], [SIDR]…
  - Pros: Global view, optimal routing path, light configuration, no routing calculation on network node (typically, but variants available)
  - Cons: Concerns on scalability, out of band control network introducing another source of failure, and inefficient exception handlings as controller involved

Any good way to combine the advantages of both to achieve the enhanced scalability, improved resilience and optimized performance?

Emerging trend: Hybrid routing (one example)



Primus: Fast, Scalable and Robust Centralized Routing for Data Center Networks

Infocom 2021

REF: https://dl.acm.org/doi/pdf/10.1109/TNET.2023.3259541

- **Targeting problems**
  - BGP as a de facto DCN routing protocol has open issues, such as routing convergence, hard to control and manage the network.
  - Centralized routing has the drawback of long status distributing procedure, as well as gracefully handing of control plane failover.
- **Key designs of Primus**
  - Centralized master monitors all the link-states and each switch calculates routes by itself
    - Stateless controller collects/disseminates all the LSes.
    - Switch reports its link status to the controller.
    - Switch does the routing path computation and update based on **preinstalled** topology and rules, as well as the LSes.

Inspiration and potential work: Hybrid routing for computing cluster network

- Different approaches for hybrid routing from research, mostly include
  - Proper division of routing functions between the centralized controller and individual network nodes
  - Fully use of the prior knowledge (topology, path, traffic pattern) and modularity of the computing cluster network
    - Regular address allocation according the structured modularity of the clusters
    - Pre-calculation of routing path based on the known topology and traffic pattern
    - Predefined and pre-deployed rules for fast route convergence to reduce per-node calculation overhead
- Next steps:
  - Revise the current draft to elaborate the commonalities of hybrid routing from research
  - Prepare a new draft to cover the overall architecture of hybrid routing and potential future work in IETF.
  - Welcome to join (liyizhou@huawei.com)