VL2: A Scalable and Flexible Data Center Network

Presented by: Wenqing & Yanmeng

Cloud-service Data Center

Tenets:

- Agility: assign any servers to any services to boost utilization
- Scaled-out ability: use large pools of commodities to achieve performance, availability, and low cost

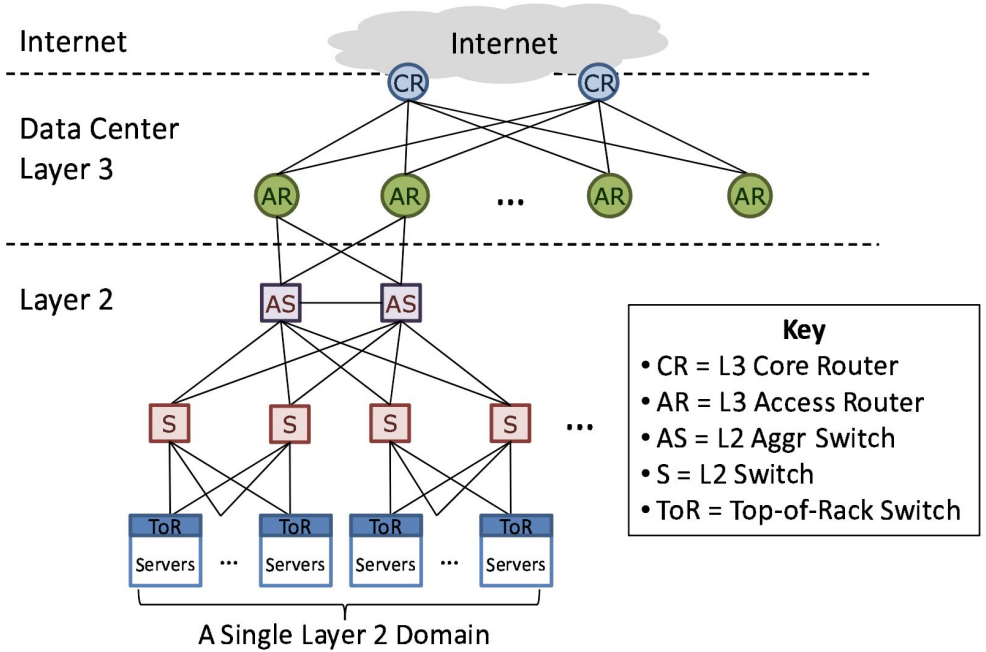
What is VL2?

- The first data center network that enables agility using scaled-out topology

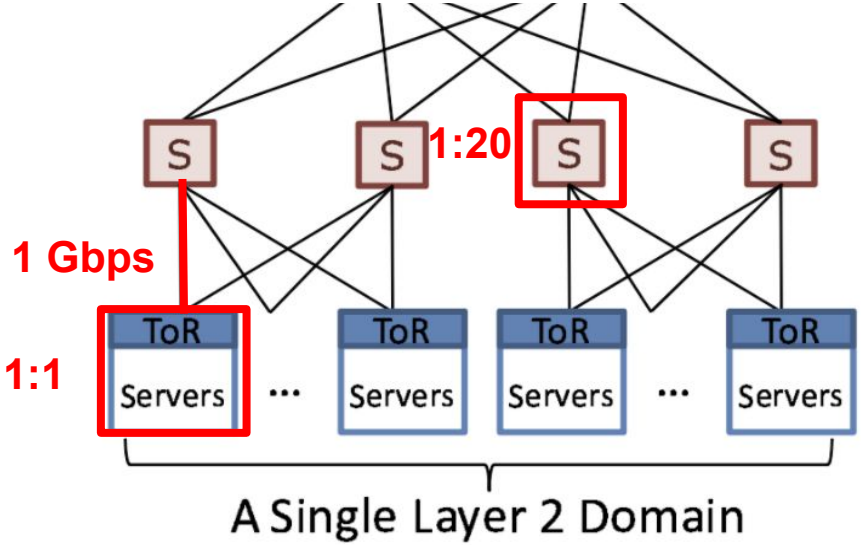
Conventional Data Center Network Architecture

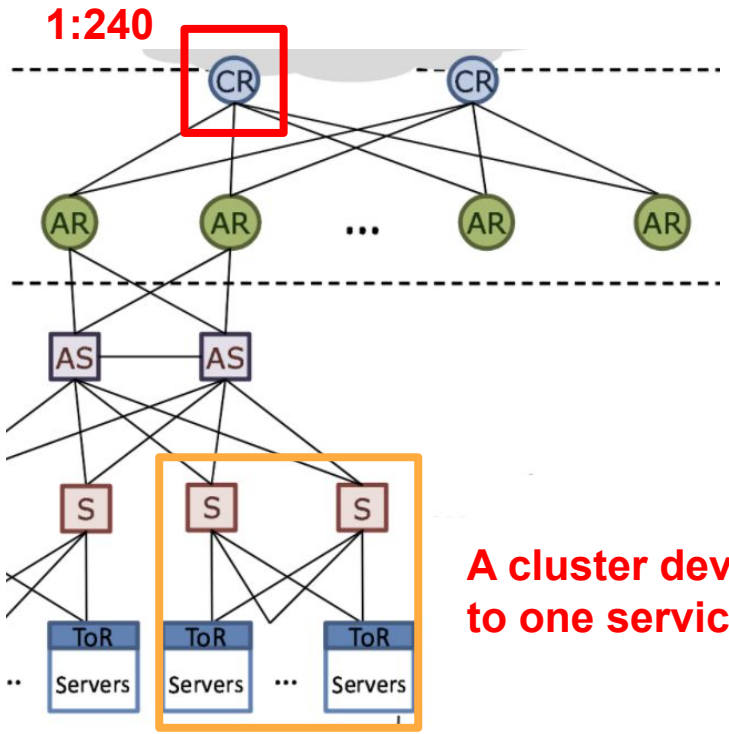
Drawbacks:

- Limited server-to-server capacity
- Fragmentation of resources
- Poor reliability & utilization



High over-subscription ratio & Resource Fragmentation



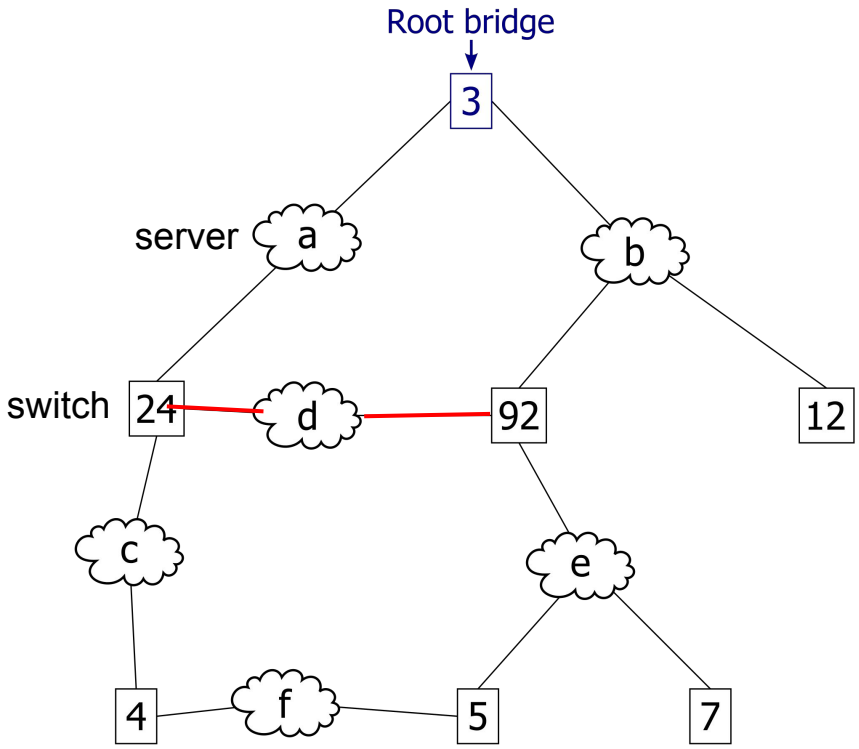


A cluster devoted to one service

Over-subscription ratio: The ratio of a network's maximum potential demand to its full rate.

Poor reliability & utilization

- 1:1 redundancy: 50% servers & links used to account for failure
- Spanning Tree Protocol
  - Always select lowest-cost path
  - Offers at most 2 paths



Spanning Tree Protocol

Source: https://commons.wikimedia.org/w/index.php?curid=3503391

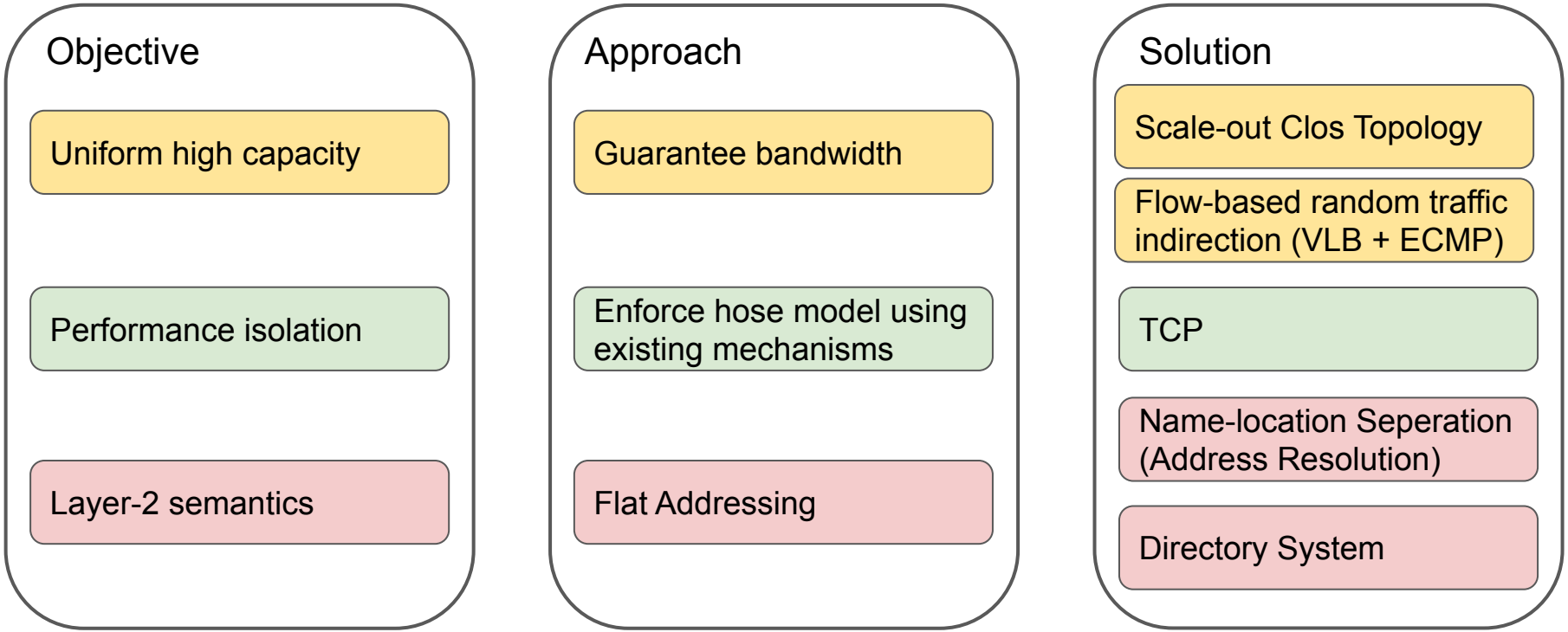
Measurements

- Data-Center traffic analysis
  - Traffic volume between servers : traffic entering/leaving data center = 4:1
  - Bandwidth demands: between servers > external host
  - Network is the bottleneck of computation
- Traffic matrix analysis
  - Volatility: hard to summarize & predict traffic patterns
- Flow distribution analysis
  - >99% flows are <100MB
  - Rare to see >GB flows
  - On average, >50% time, ~10 concurrent flows/machine; 5% time, ~80 flows/machine
- Failure characteristics
  - Most failures are small in size: 95%> failures involve <20 devices
  - Downtime can be significant
  - No obvious way to eliminate failures from top of the hierarchy

Implications

- Focus more on traffic management between servers
- Avoid multi-layer tree structure
  - For higher utilization
  - For better failure resilience
- Need mechanisms to cope with traffic unpredictability and volatility
- Randomizing path selection at flow granularity will not cause perpetual congestion if unlucky flow placement happens

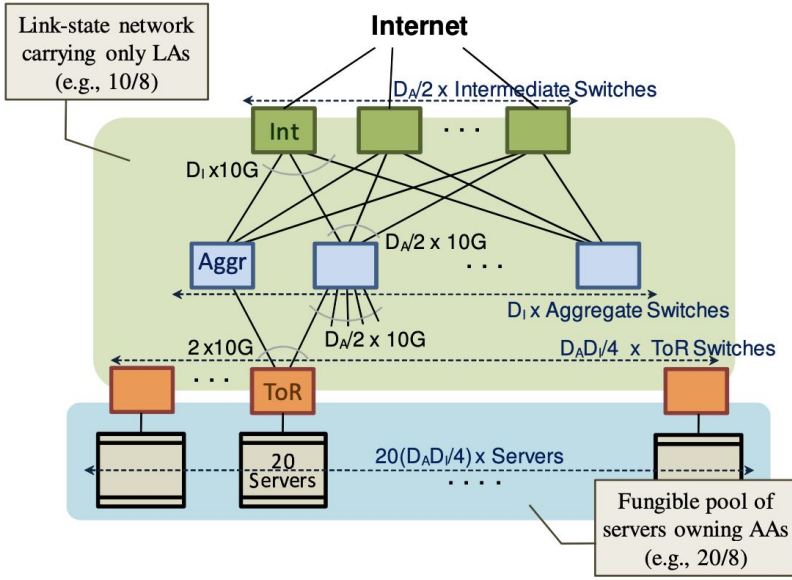
VL2 Design Principles



Hose Model: the ingress/egress bandwidth of each node is constrained.

Scale-out Clos Topology + Valiant Load-balancing (VLB)

- Equal Cost Multi-path Forwarding (ECMP) + IP Anycast
- Links between the Intermediate switches (IS) and the Aggregation switches (AS) form a complete bipartite graph
  - Provide huge aggregation capacities and extensive path diversity
  - Provide huge bisection bandwidth
- Routing is resilient: need a random path connecting a from TOR and to an IS
- Ensure robustness to failures



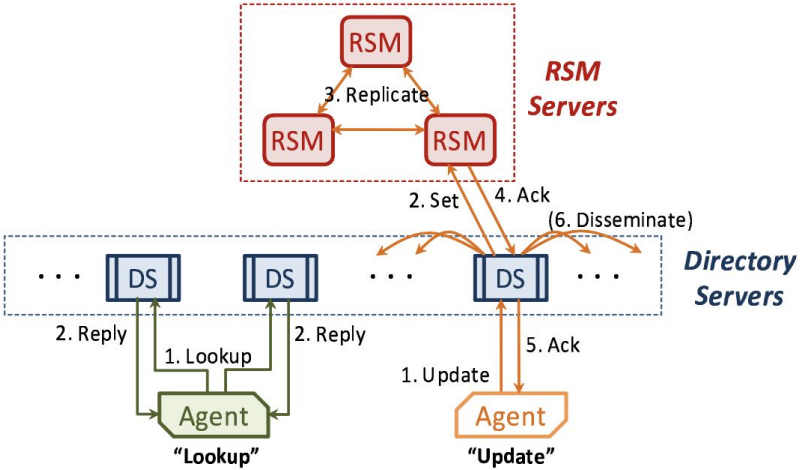
Address Resolution

Name/Location Separation

- Switches run link-state routing and maintain only switch-level topology
  - Cope with host churns with little overhead
- Enable Agility with low-cost switches
- Allow VM migration
- Provide access control

Directory System

- Quick lookups with reactive cache support
- Update with reliability
- Eventual Consistency



Evaluation

Uniform high capacity

- All-to-all data shuffle traffic matrix:
  - 75 servers
  - Each delivers 500MB
- Completes in 395s
- Aggregate goodput: 58.8Gbps, Maximal achievable goodput: 62.3Gbps
  - 10 times better
- Network efficiency 58.8/62.3 = 94%

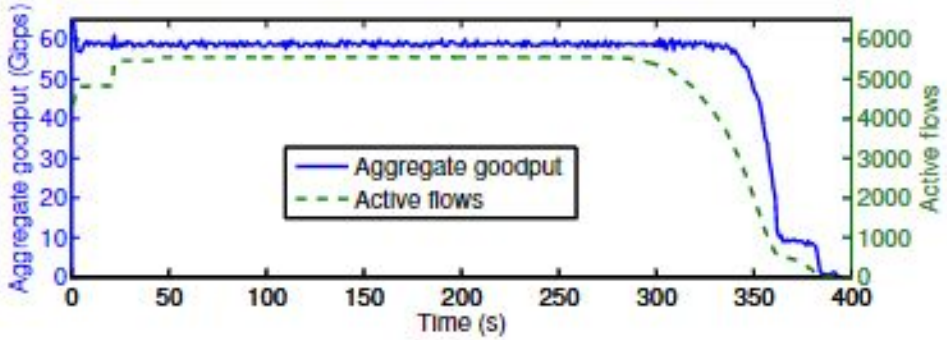


Figure 9: Aggregate goodput during a 2.7TB shuffle among 75 servers.

Evaluation

VLB Fairness

- 75 node testbed
- VLB split ratio fairness index
- Averages > 0.98%

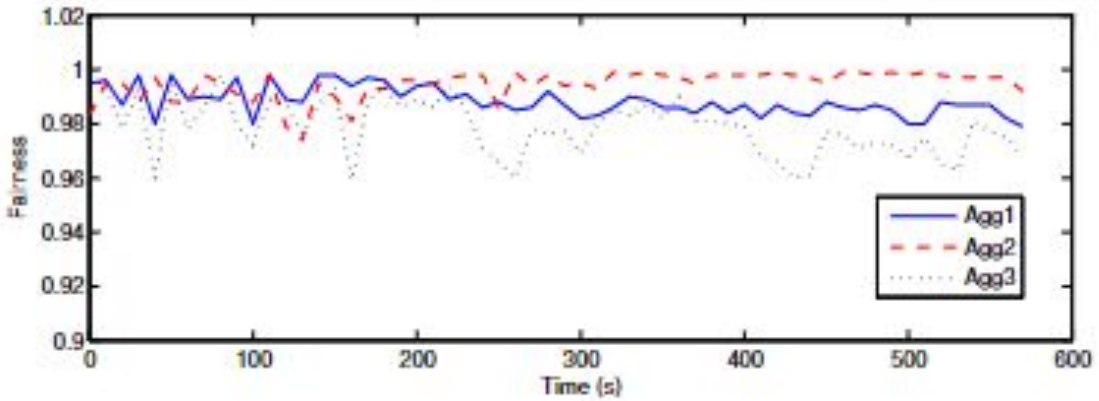


Figure 10: Fairness measures how evenly flows are split to intermediate switches from aggregation switches.

Evaluation

Performance Isolation

- Service 1:
  - 18 servers do single TCP transfer to another server
  - Starting at time 0 and lasting throughout the experiment
- Service 2:
  - Start one server at 60s and assign a new server every 2s for a total of 19 servers
  - Each one starts a 8GB transfer over TCP as soon as it starts up

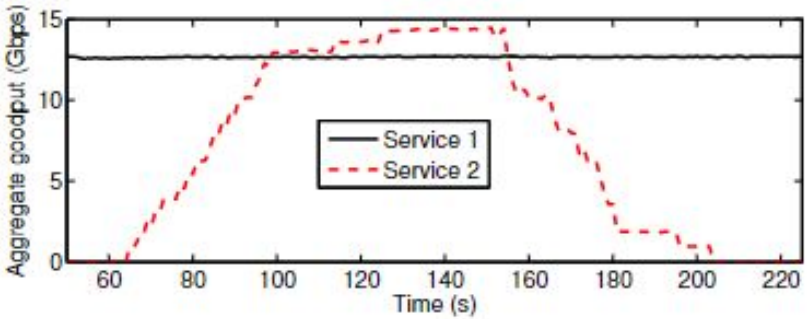


Figure 11: Aggregate goodput of two services with servers intermingled on the ToRs. Service one's goodput is unaffected as service two ramps traffic up and down.

No perceptible change in Service 1

Evaluation

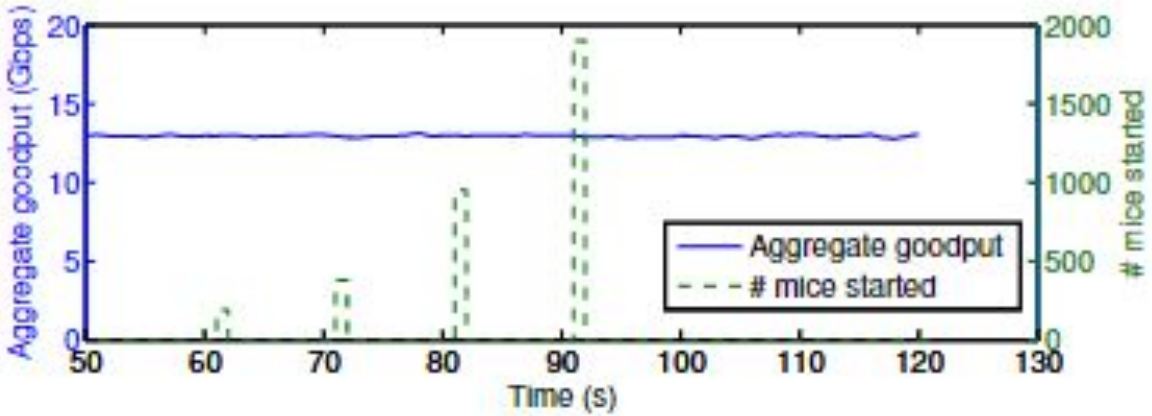


Figure 12: Aggregate goodput of service one as service two creates bursts containing successively more short TCP connections.

No perceptible change in Service 1

Evaluation

Convergence After Link Failures

- Max capacity degrades gracefully
- Restoration is delayed
- Restoration does not interfere with traffic and the aggregate throughput eventually returns to its initial level

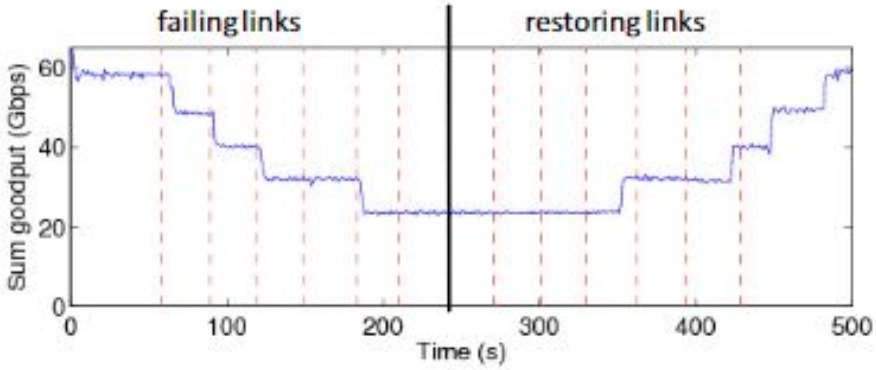


Figure 13: Aggregate goodput as all links to switches Intermediate1 and Intermediate2 are unplugged in succession and then reconnected in succession. Approximate times of link manipulation marked with vertical lines. Network re-converges in < 1s after each failure and demonstrates graceful degradation.

Downsides

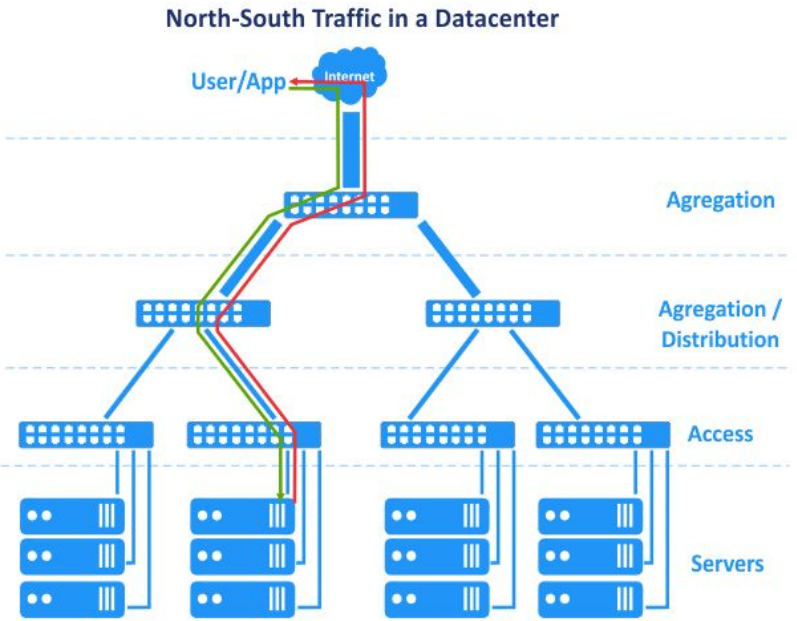
- Extra servers are need to support the VL2 directory system
- All links and switches are working all the time
  - Not power efficient
- Evaluation of real time performance is missing
- They only looked at “a highly utilized 1,500 node cluster in a data center that supports data mining on petabytes of data”

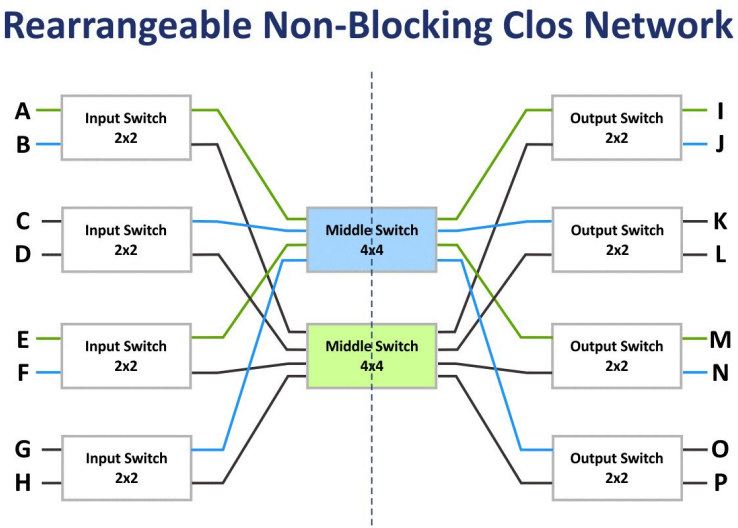
Discussion

- How will you change VL2 if traffic patterns were predictable/can be modeled really well by some learning algorithm?
- How would you implement such a change and how does it compare with the implementation in the paper?
- What optimizations can be performed in VL2 for example in the topology, network devices, etc using new hardware/software available today?
- What topology to use in today's data centers?

Discussion Doc: https://tinyurl.com/cse550vl2

Some topologies





Source: https://www.nakivo.com/blog/msp-network-topology-for-beginners/

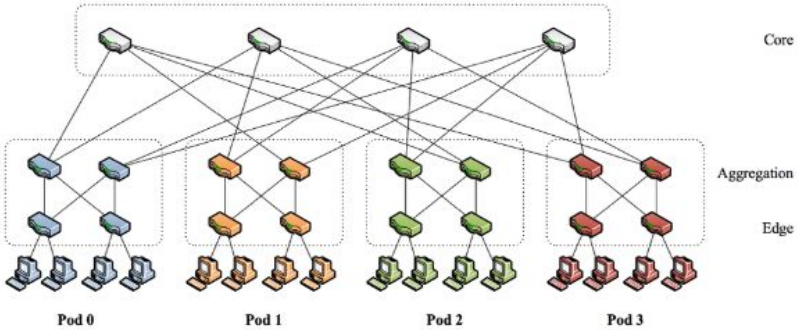
Related Work

- SEATTLE
  - A Scalable Ethernet Architecture for Large Enterprises
- PortLand
  - A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric
- BCube
  - A High-Performance, Server-centric Network Architecture for Modular Data Centers

Related Work

PortLand

- VM migration
- Administrator do not need to configure switches
- Any end host should efficiently communicate with others
- No forwarding loop
- Failure detection



VL2

- Support huge data centers with uniform high capacity between servers
- Performance isolation between servers
- Ethernet layer 2 semantic

References

- VL2: A Scalable and Flexible Data Center Network
- SEATTLE: A Scalable Ethernet Architecture for Large Enterprises
- PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric
- BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers
- Routing & Architecture Comparison about: DCell, PortLand, VL2, BCube, MDCube