OPENFABRICS ALLIANCE

2021 OFA Virtual Workshop

Infiniband reliability engineering - stories from a public cloud

Vladimir Chukov, Sr. Network Engineer

1&1 IONOS SE

IONOS by 1&1

INTRODUCTION

- **Reliability Knowledge base**
- **Failures and mitigation**
  - Physical
  - Link (+Subnet Management)
  - Network
  - Transport
  - ULP

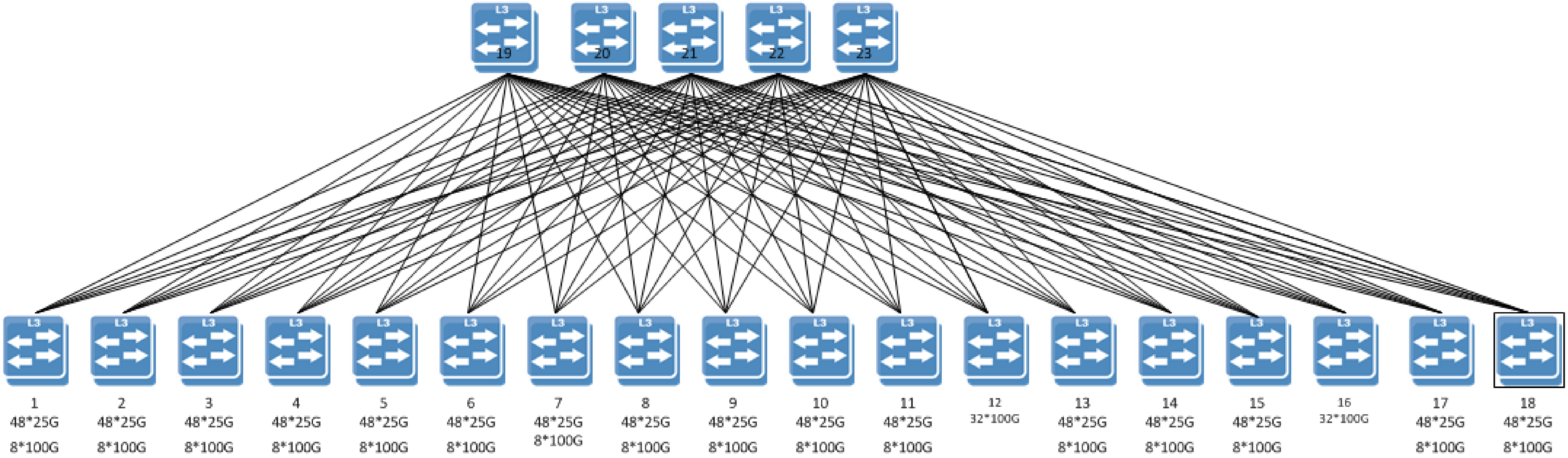
RELIABILITY KNOWLEDGE BASE

- **Vendor ecosystem comparison**
  - Guides/solutions/training and certification programs
  - PRM/Architecture Specifications and rare articles, mailing lists
- **Cloud, Finance and some HPC fabrics do not tolerate downtime**
- **IETF, RIPE and smaller communities that address design and availability, from protocols to users -> on the way there as IB becomes a commodity**
- **Books - from engineers to engineers -> we need those**
- **OFA presentations help**

FAILURES AND MITIGATION

- **Survivorship bias + collection of assumptions + Ethernet influence**
- **Layer 1**
  - DACs/Optical are tested well - BER rate on par/even better
  - Failure rates low
  - Errors are easy to catch with the counters
  - Chassis do not maintain signal integrity in all combinations (QDR/FDR)
  - Colocation personnel training takes time (all over the world)
  - Avoid recabling

Example of a modern Ethernet Fabric



FAILURES AND MITIGATION

- **Layer 2**
  - Modern fabrics are L3 based, dynamic routing
  - No L2 multicast in modern DC designs
  - Point to point links
  - ECMP
  - No Auto-Negotiation
  - No chassis/dual supervisors/kernel sync
  - Complexity is pushed up the stack into overlays, resulting in significant stability increase
  - Handshake/keepalive-based fabric compared to centrally programmed LFTs
  - Simplicity in DCN (D.Dutt)

FAILURES AND MITIGATION

- **Layer 2**
  - Constant growth means more variety, compared to largely static HPC fabrics
  - Switch firmware - could be catastrophic for the fabric.
    - Propagation of errors in the fabric
    - Managed switches with local flashing
      - Switchdev/SONiC would be great
    - Externally managed
      - Updates before the installation (scale issues)
  - Stay within a generation, max 2 generations in a cluster
    - The smallest penalty is negotiation errors

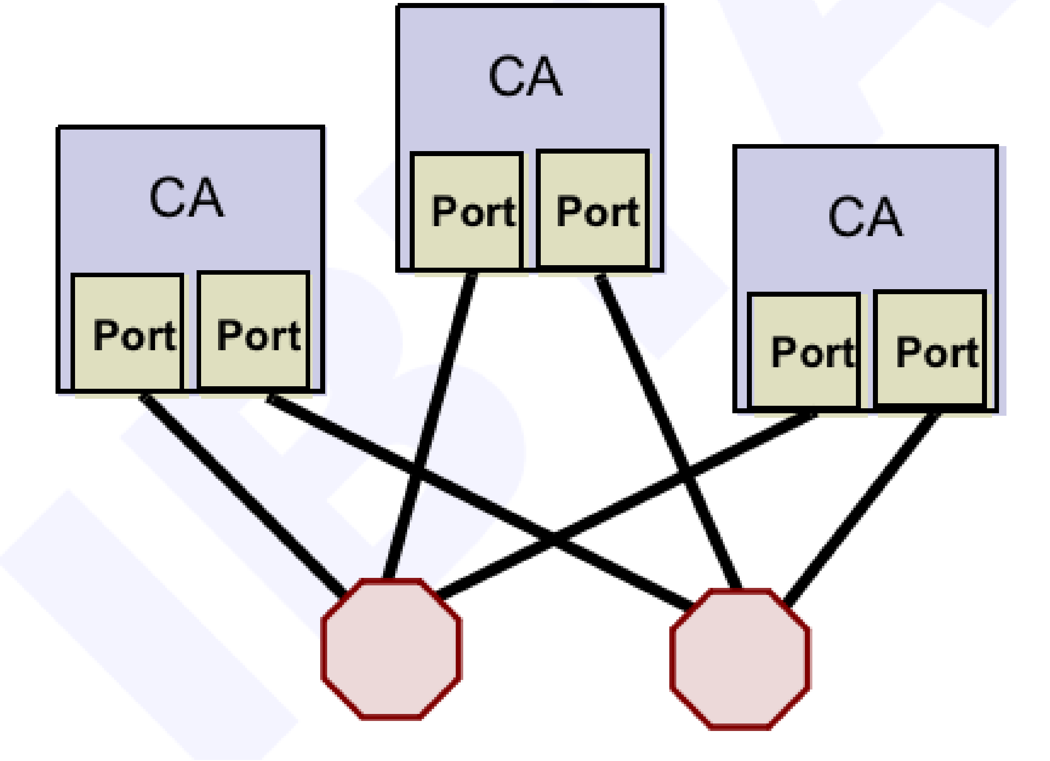
FAILURES AND MITIGATION

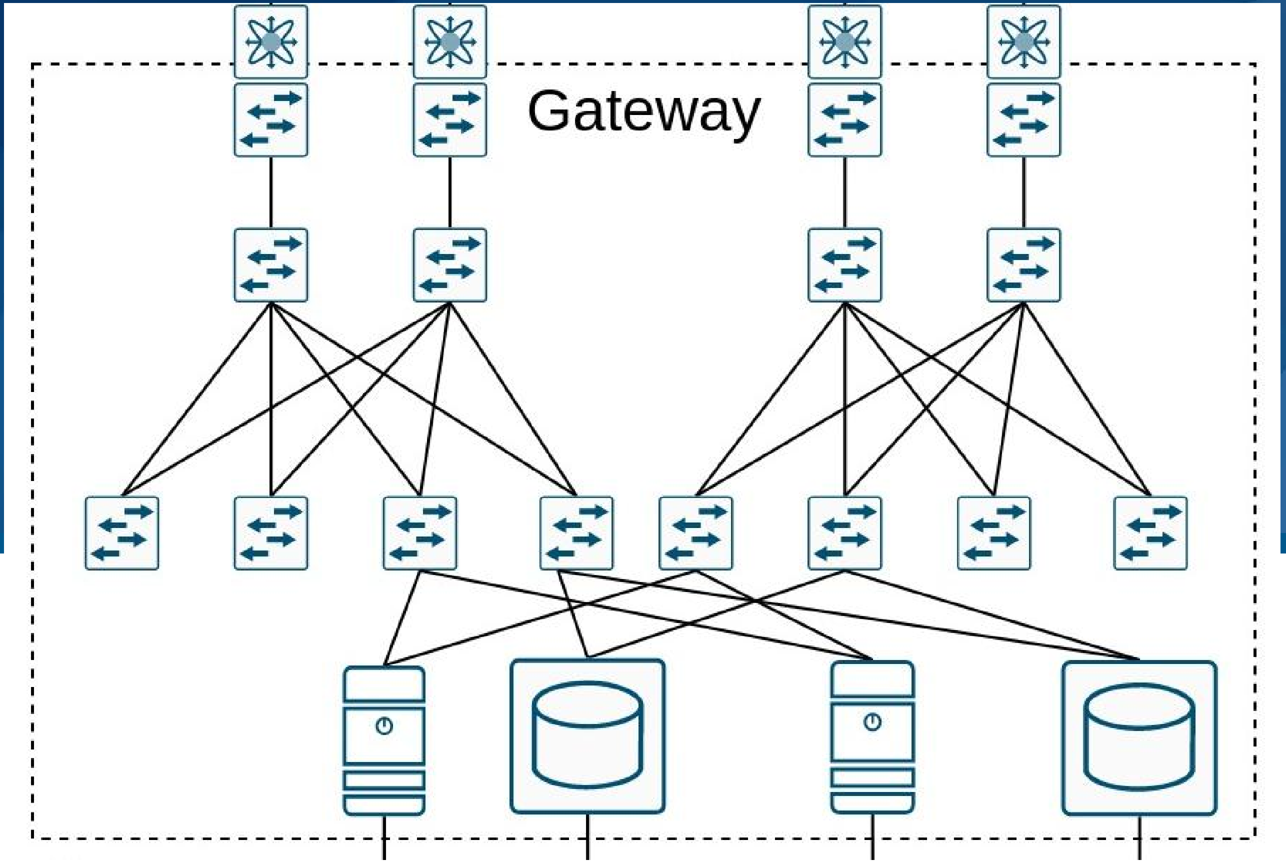
- **Layer 2 (cont.)**
  - 70% of the incidents come from changes (Google SRE book)
  - Centralized management with SMs
  - Running a few (not too many) SMs is better
  - However, there were events where the first SM hung, 2nd took over and then it hung too

  - 0x02 -> log_notice: Reporting Urgent Notice "Link state change" from switch LID 67, GUID 0xb8599f03009cda80
  - 0x01 -> req_determine_mkey: ERR 1107: Outgoing physp is null on non-hop_0!
  - 0x01 -> log_rcv_cb_error: ERR 3111: Received MAD with error status = 0xC SubnGetResp(SMInfo), attr_mod 0x0, TID 0x2a3b536

FAILURES AND MITIGATION

- **Layer 2 (cont.)**
  - SM HA with OOB sync is there, but depends on the SMs to never get isolated
  - Other advanced SM HA features (guid2lid, SA DB sync) add more complexity, but do not guarantee availability
  - Limited CPU resources on the switches
    - MC joins/leaves
  - Heavy sweeps + ucast-cache to ensure unhealthy nodes are removed
    - Possibly stuck firmware + kdump
    - Fabric can be flooded
    - Watchdog on every server could work as a mitigation


CA
Port
Port
CA
Port
Port
CA
Port
Port


Gateway

FAILURES AND MITIGATION

- **Layer 2 (cont.)**
  - Way out: Dual fabric approach (IBTA Vol.1) p.1190
  - Human error and linking fabrics
    - LIDs reassigned + MC groups recreated
    - SM/M_key, protection level 2 as a countermeasure
  - Not a waste of resources
    - ECMP for network and storage traffic
    - Blazing fast rollouts
  - Clos-(3,5) - most tested, well-known and feature-rich topology
    - Rings lead to issues

FAILURES AND MITIGATION

- **Layer 3**
  - Out of scope
- **Layer 4**
  - Mostly out of scope
  - UD is recommended today for IPoIB (multiqueue)
    - 4096 is the max available MTU
    - RC wasn't super helpful

FAILURES AND MITIGATION

- **Layer 5 (IPoIB)**
  - On top of hardware-replicated MC groups
  - SM will clean/recreate those groups on start or failover
  - At least 40ms outage for a < 10 switches fabric
  - Failures are hard to detect
    - Routing protocols (OSPF, BGP) on top of IPoIB
    - iproute2 won’t help much
    - ibdump

FAILURES AND MITIGATION

- **Layer 5 (IPoIB)**
  - small test base -> more likely to run into corner cases
  - especially for IPv6
  - An example of an IPoIB bug (pkeys removed)
  - Shared fate/No redundancy for software components (mlx4/mlx5, ib_ipoib) -> mitigated with testing

OPENFABRICS ALLIANCE

2021 OFA Virtual Workshop

THANK YOU

Vladimir Chukov, Sr. Network Engineer

vladimir.chukov@cloud.ionos.com,

linkedin.com/in/vchukov

1&1 IONOS SE

IONOS by 1&1