ZooRoute: Enhancing Cloud-Scale Network Reliability via Candidate Path Provisioning and Overlay Proactive Rerouting

Xiaoqing Sun, Xing Li, Xionglie Wei, Tian Pan, Ju Zhang, Bowen Yang, Yi Wang, Ye Yang, Yu Qi, Le Yu, Chenhao Jia, Zhanlong Zhang, Xinyu Chen, Jianyuan Lu, Shize Zhang, Enge Song, Yang Song, Rong Wen, Biao Lyu , Yang Xu, Shunmin Zhu

alibaba_cloud_network@alibaba-inc.com

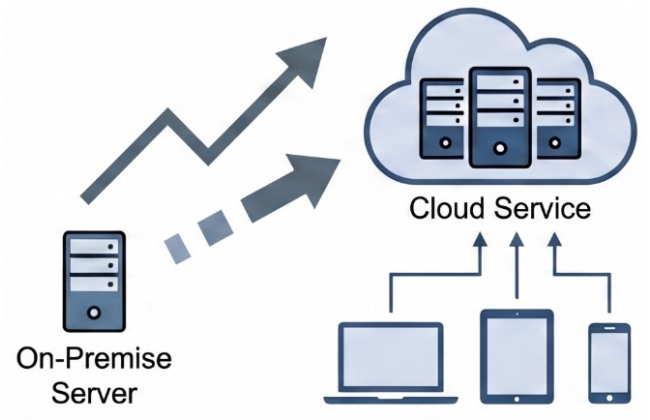
Alibaba Cloud

浙江大學 ZHEJIANG UNIVERSITY

復旦大學 FUDAN UNIVERSITY

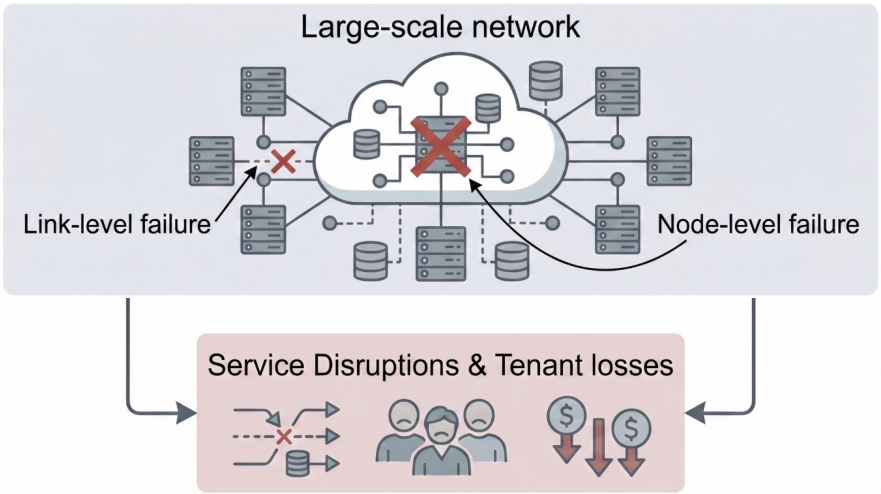
Cloud Network Reliability

High growth in "Cloud Shift"



Gartner reports: IT spending on public cloud services is expected to exceed $1 trillion by 2017.

Inevitable network failures



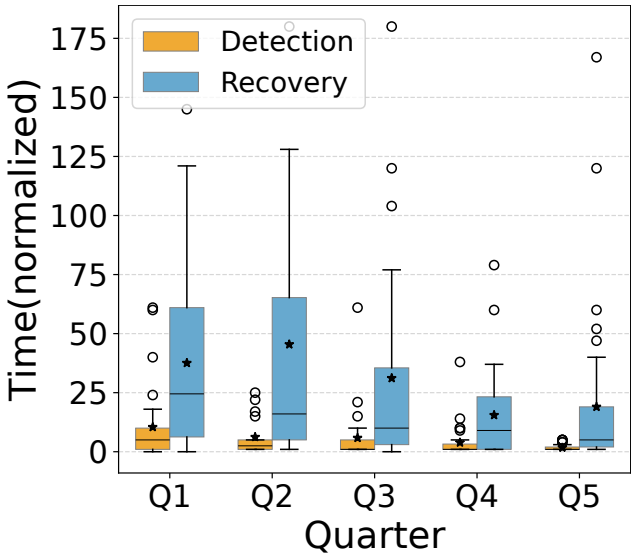
Network reliability has become a key concern

Detection vs. Recovery Gap

Statistical results of 150+ network failures in Alibaba Cloud from 2022-01-01 to 2023-03-31 shows:

We can detect failures quickly, but cannot ensure timely recovery with existing telemetry solutions.

- Average recovery time is ***3-10x*** detection time
- Only ***14%*** failures resolved promptly
- ***~20%*** failures have recovery time ***16x*** detection time



We need **fast, deterministic** failure bypass!

Existing Solutions & Limitations

Application layer

Active/standby structure

Multi-path protocols

- Extra CapEx and OpEx;
- Uncoordinated with CSP's infra, may amplify failures.

Physical layer

Traffic engineering

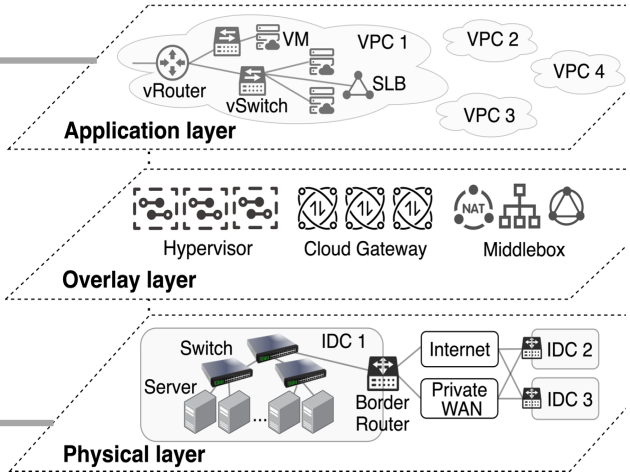
Global reconvergence takes 10+ *min*; High underlay impact.

Fast Reroute

Limited capacity; Cannot handle unexpected failures

Protective Reroute

Opportunistic; Requires IPv6 FlowLabel hashing



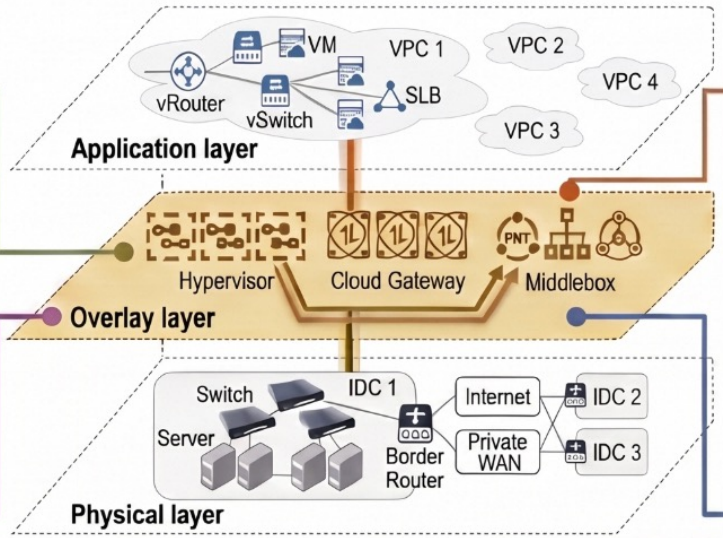
ZooRoute Design Goal

Tenant-Transparent

Cloud-native failure recovery as a service; no modifications to tenant workloads or traffic.

Fast & Deterministic

Proactive probing & validation of candidate paths; one-shot switch to guaran healthy path.



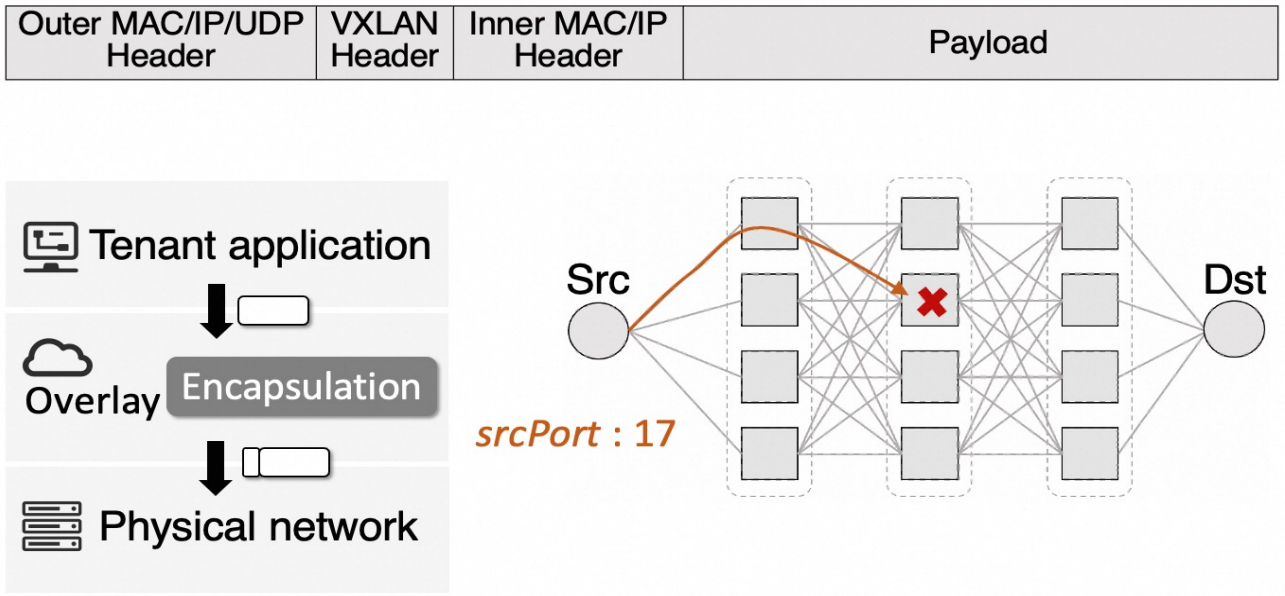
Underlay-Agnostic

Exploits overlay path diversity; zero changes required to existing underlay devices.

Unified Framework

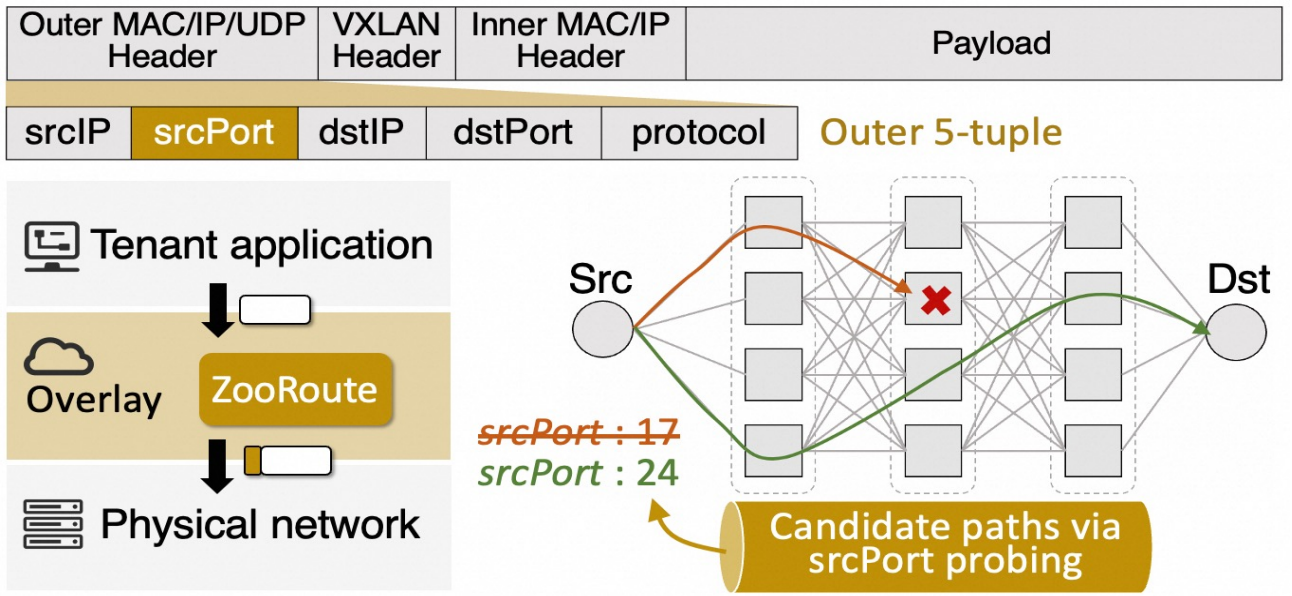
One recovery mechanism for all scenarios: VM-VM, VM-Internet, VM-IDC, intra & inter-region.

ZooRoute Insight



Manipulate VXLAN **outer *source port*** at overlay layer to influence ECMP hashing

ZooRoute Insight: Source Port-Based Path Switching



Step 1: Path Probing

Probe paths with varying source ports between VTEPs

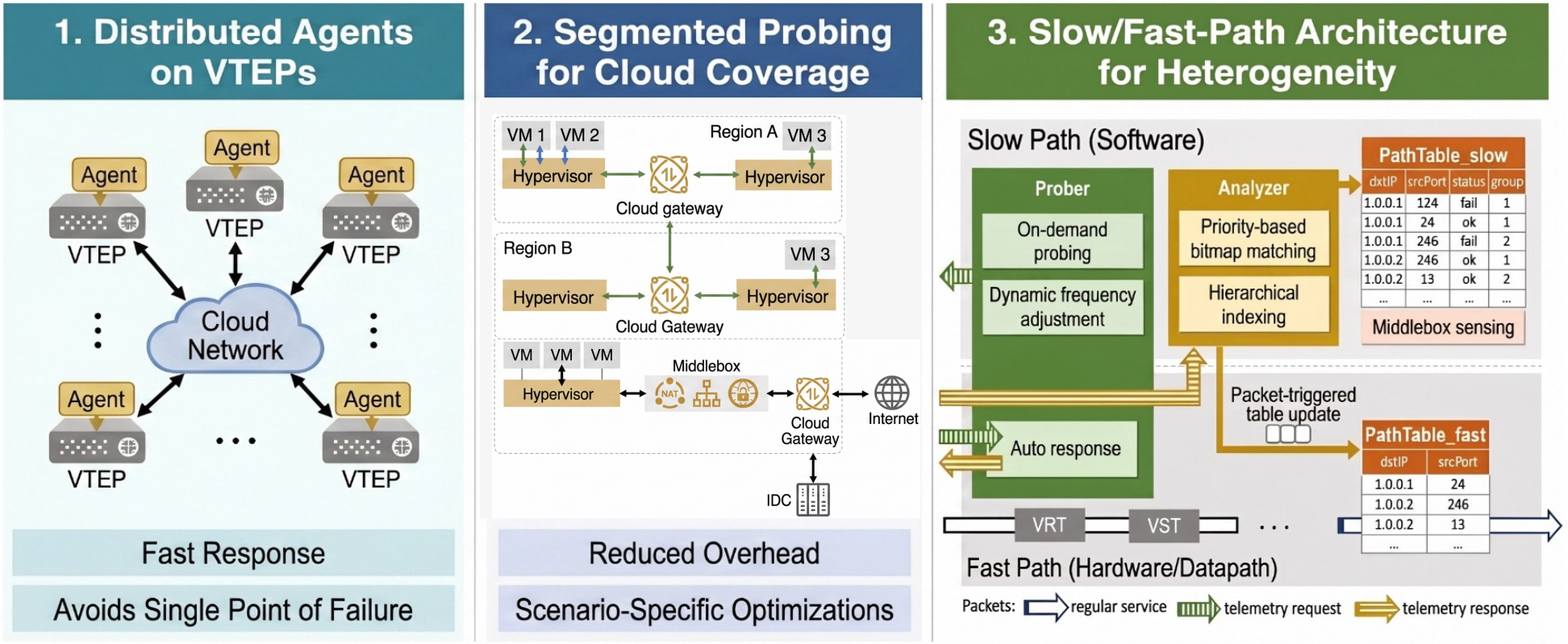
Step 2: Path Recording

Maintain healthy candidate paths in compressed path table

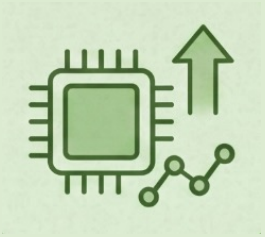
Step 3: Path Switching

Switch to a pre-validated healthy path upon failure

ZooRoute Architecture

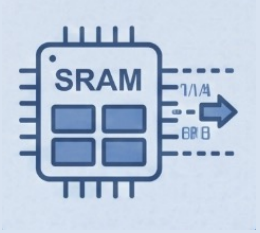


Challenges in Large-scale Deployments



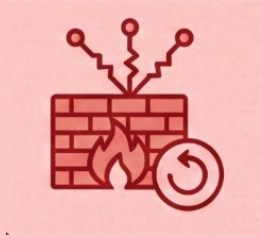
C1: VTEP CPU Overhead of Path Probing

Full-mesh probing at cloud scale is prohibitively expensive. Alibaba Cloud regions can have 200k+ servers. Need to maximize network visibility within a strict cost budget.



C2: Path Recording on Tofino Gateways

Tofino has limited on-chip SRAM (~10MB). Gateways must track source port status for 50k+ VTEPs. PCIe updates are too slow (10k TPS) for massive failure scenarios.



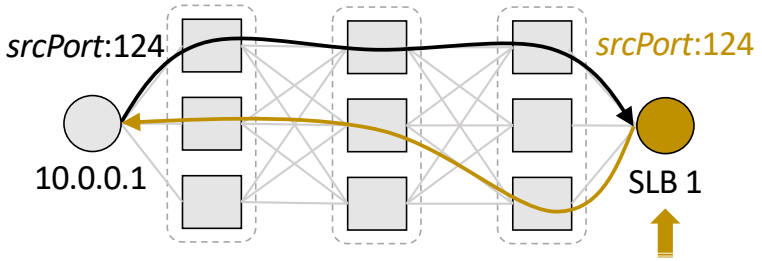
C3: Path Switching at Stateful Middleboxes

Changing source ports may rehash flows to different middlebox instances (e.g., SLB, NAT), breaking session affinity and triggering connection resets.

Path Probing: On-Demand & Frequency Adjustment

On-Demand Probing

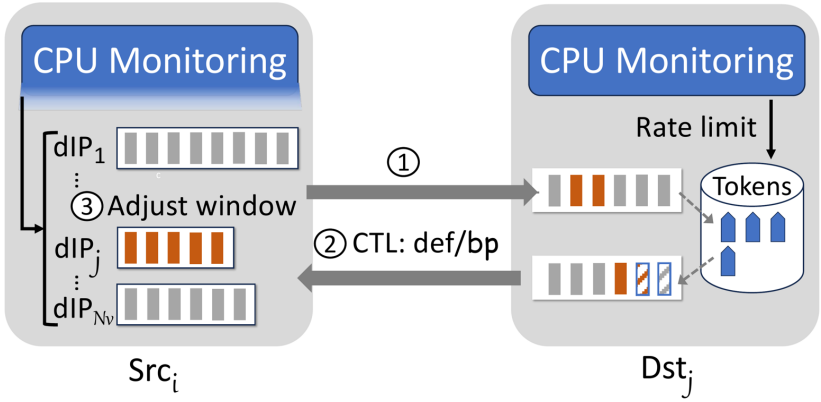
- **Active VTEP probing:** Probe only when tenant traffic exists (99% VTEPs talk to <25% peers)
- **Active source port probing:** Full-probe (Np=32) vs. Partial-probe (active ports only)
- **Source port learning:** Stateful middleboxes learn ports from session tables passively



Session Table				
Outer srcIP	Outer srcPort	Inner dstIP	Inner dstport	...
10.0.0.1	124	192.0.0.3	8001	...
10.0.0.2	20	192.0.0.1	8080	...

Dynamic Frequency Adjustment

- Monitor CPU utilization per VTEP
- Exponential backoff when CPU > 80%
- Destination-side token bucket rate limiting
- CTL field for backpressure signaling

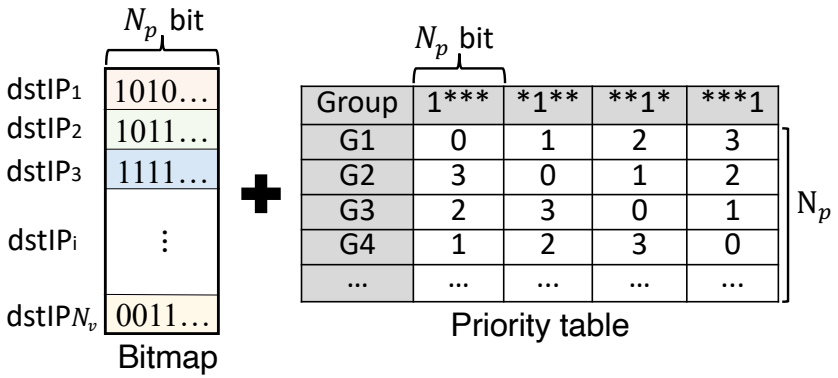


***Over 95%* probing cost reduction**

Path Recording: Table Compression

Priority-Based Bitmap Matching

For cross-region G-G scenarios on Tofino gateways

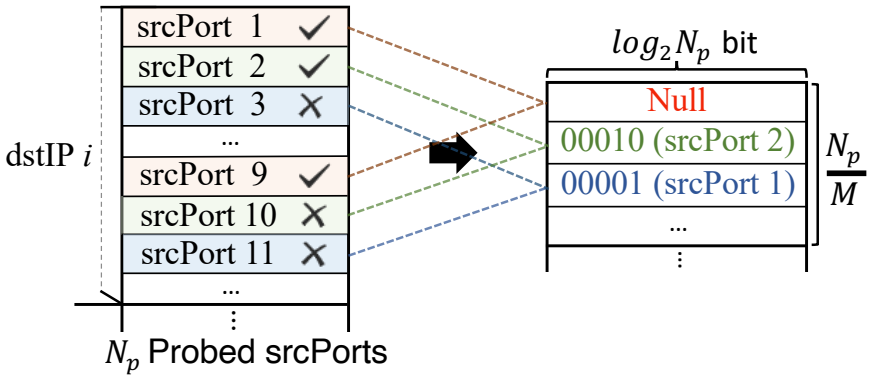


- Bitmap per destination IP: "1" = available
- Priority groups with circular-shift masks
- Inner 5-tuple hash selects group

***~15x* memory compression**

Hierarchical Indexing

For intra-region H-G scenarios (50k+ VTEPs)

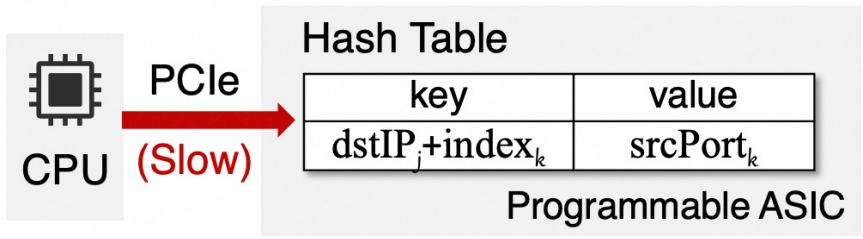


- Group every M source ports into one entry
- Empty entry: all ports OK, randomly pick one
- Filled entry: use recorded working port

***~25x* memory compression**

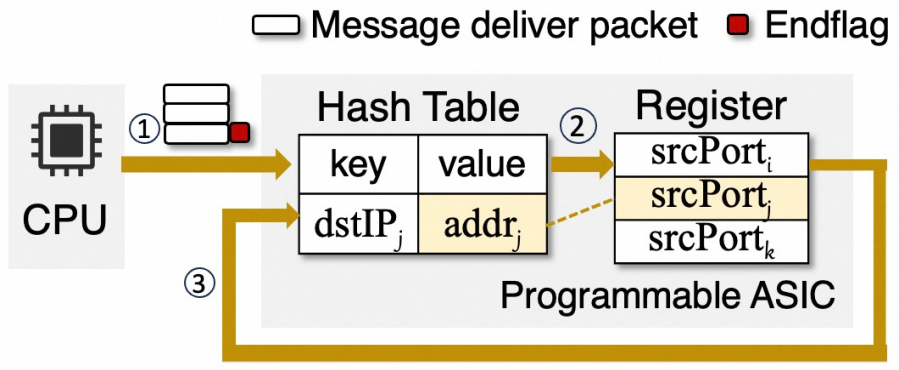
Path Recording: Packet-Triggered Updates

Problem: PCIe Bottleneck



- PCIe max throughput: ~10k Tps
- For 1.6M updates (50k VTEPs × 32 ports):
  ~ 3 minutes

In-Network Updates



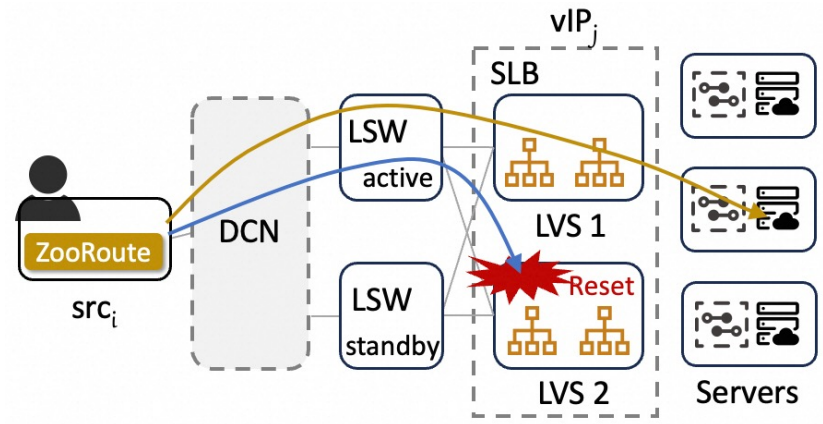
- CPU sends negotiation packets with PKT markers.
- Tofino extracts fields & updates registers **at line rate.**
- Batched updates via loopback re-injection until EndFlag.

***~1000x* Faster than PCIe**

Path Switching: Middlebox Sensing

Problem: Connection resets

Changing source port may rehash flows to other middlebox instances, causing session loss and resets.



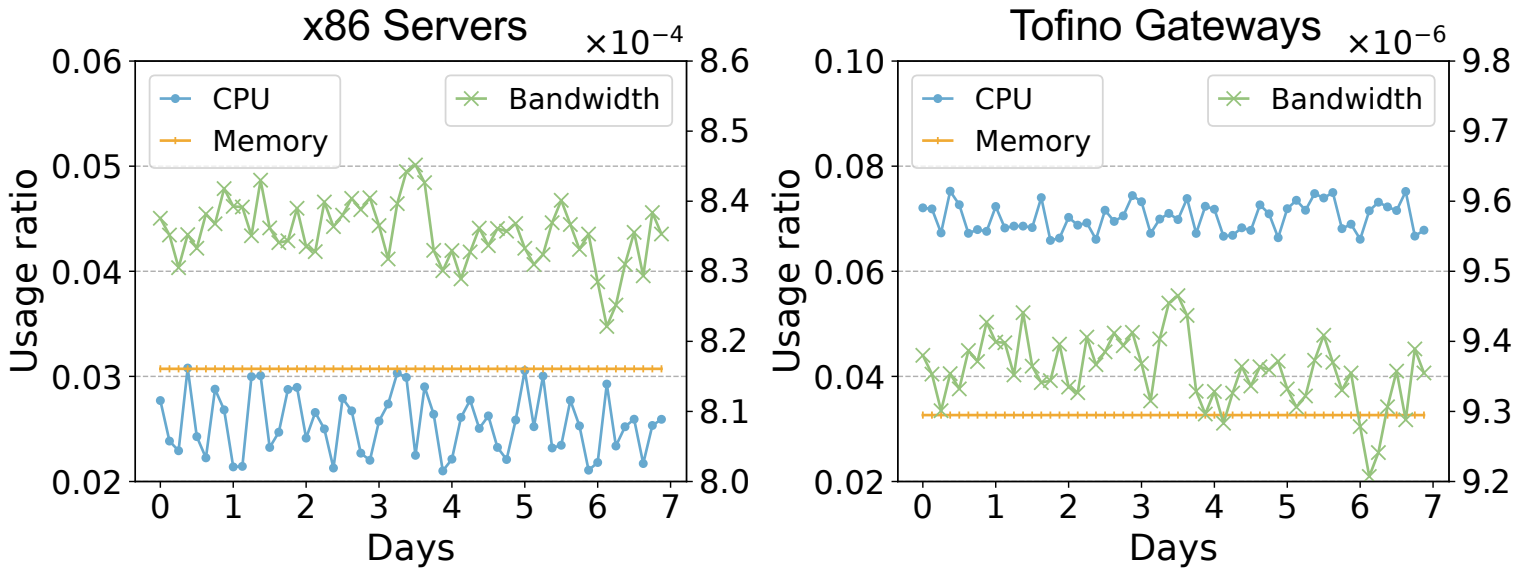
Middlebox Sensing

- Add **BackendID** field in probe packets
- Middlebox fills BackendID with MAC address
- Source VTEP groups ports by BackendID
- On failure: **prefer ports in the same group**

Path Table_slow		
dIP	srcPort	group
vIP1	24	1 (LVS1)
vIP1	50	2 (LVS2)
vIP1	51	2 (LVS2)
vIP1	124	1 (LVS1)
vIP1	12	1 (LVS1)
vIP1	135	2 (LVS2)

~ All requests succeed on *1st* try (vs. 25% failure without sensing)

Resource Overhead



x86 Server Metrics

CPU: **2-3%**
Memory: **~3%**
NIC Bandwidth: **<0.085%**

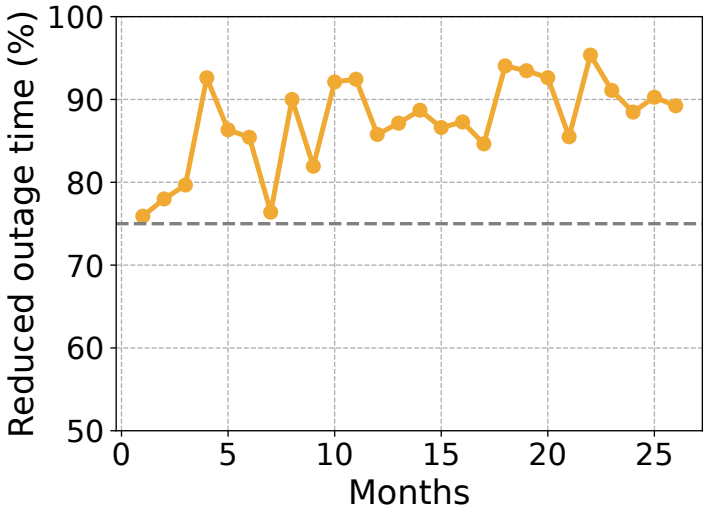
Tofino Gateway Metrics

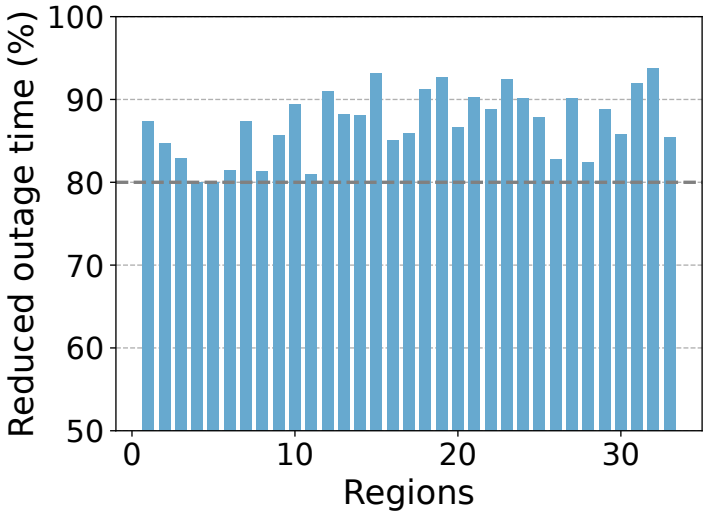
CPU: **<7.5%**
SRAM: **~3%**
Bandwidth: **<0.001%**

Acceptable ✅

Modest cost outweighs reputational & revenue risks of SLA violations.

Online Performance：26 Months in Production





93.19%
Cumulative outage time reduced

98.21%
Failures masked from tenants

33
Regions covered

500K+
Servers protected

Case Study: Top3 failure types mitigate by ZooRoute

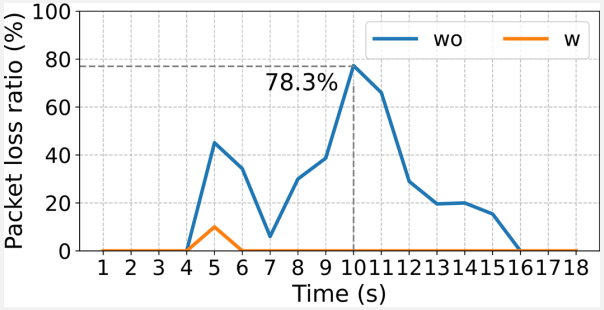
Long-Haul Link Jitter

Scene: inter-region

Avg Reduce 7.47s (84.67%)

Case1

- 12s fluctuation 78.3% peak loss
- Rerouted in <1s
- Transient loss <10%



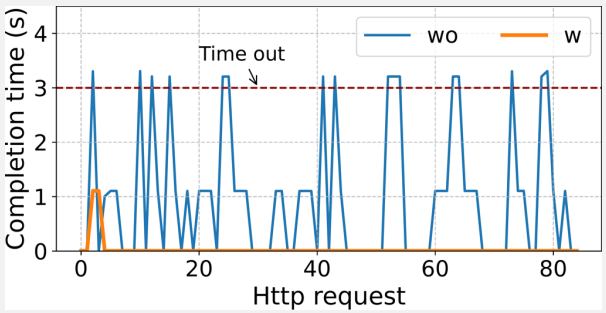
Switch Card Failure

Scene: intra-region

Avg Reduce 139.52s (93.12%)

Case2

- 2min failure
- 67% inter-AZ traffic affected
- VM w ZooRoute: short spike only



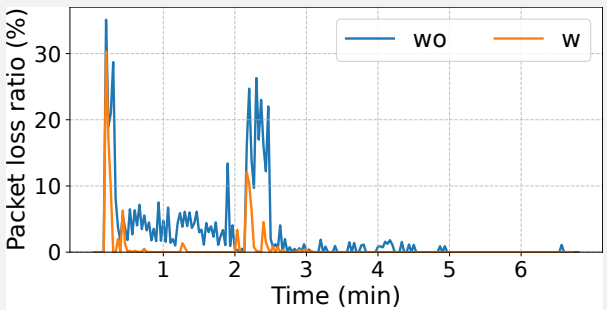
Border Routers Down

Scene: inter-region

Avg Reduce 295.33s (98.60%)

Case3

- 2 of 4 routers failed.
- Reduced loss **but full recovery limited by capacity.**



Experiences

Deterministic > Opportunistic

- **Cross-region**: avg. 79.52% ports remain healthy, worst case 24.22%.
- **Opportunistic** needs avg. 1.58 retries, worst case 17.05.
- **Deterministic** failover is crucial.

Sensitivity vs. Stability

- Mark a port “bad” only when two subsequent probes fail.
- **Hysteresis filter**: 3-out-of-5 policy.
- **Rank** candidates by failure-free duration.

Known Limitations

- **Capacity scarcity**: Underperforms with tight capacity or unstable routing states.
- **Low-rate loss**: Rare/random packet loss hard to detect with limited probes.
- **Endpoint-adjacent faults**: Path switching is less effective when the packet loss originates near endpoints.
- **VM datapath failures**: vNIC drops or queue hangs require VM-level visibility, handled by complementary system, *Zoonet*.

ZooRoute: Cloud Network Failure Recovery Service

Tenant-Transparent, Underlay-Agnostic, Within Seconds.

26 Months in Production

93.19% Outage Time Reduced

98.21% Failures Masked

Q&A

alibaba_cloud_network@alibaba-inc.com