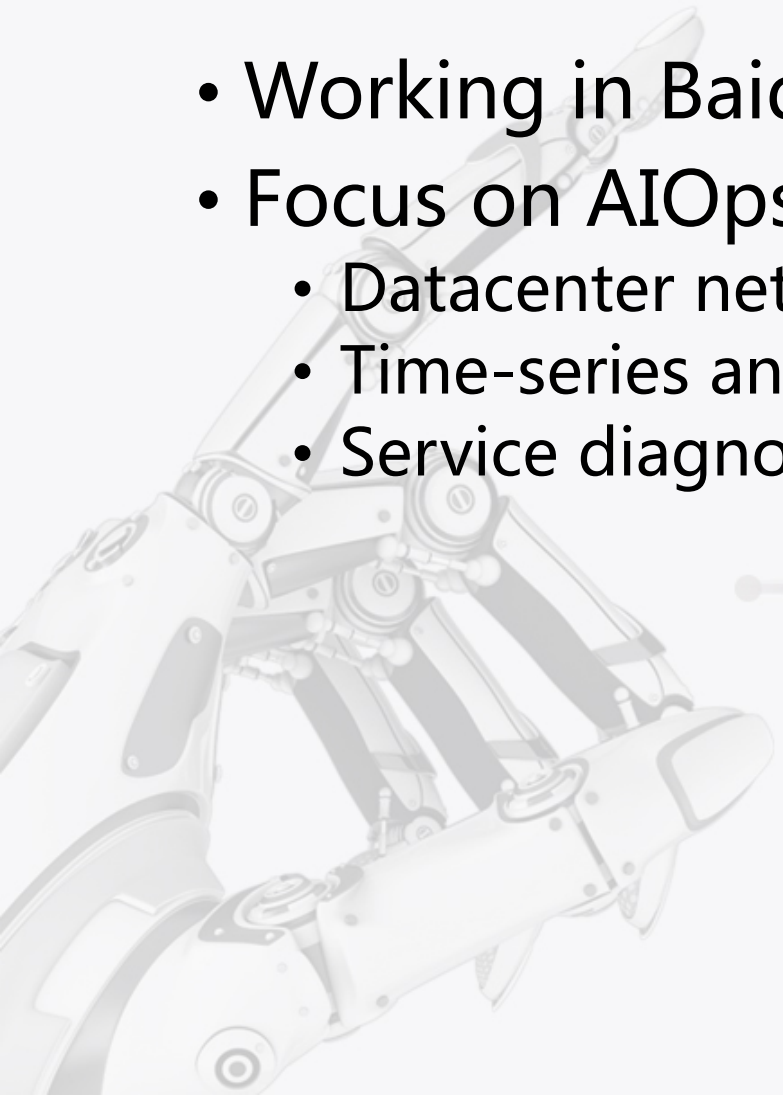


NetRadar: Monitoring the datacenter network

Yun Chen

About me

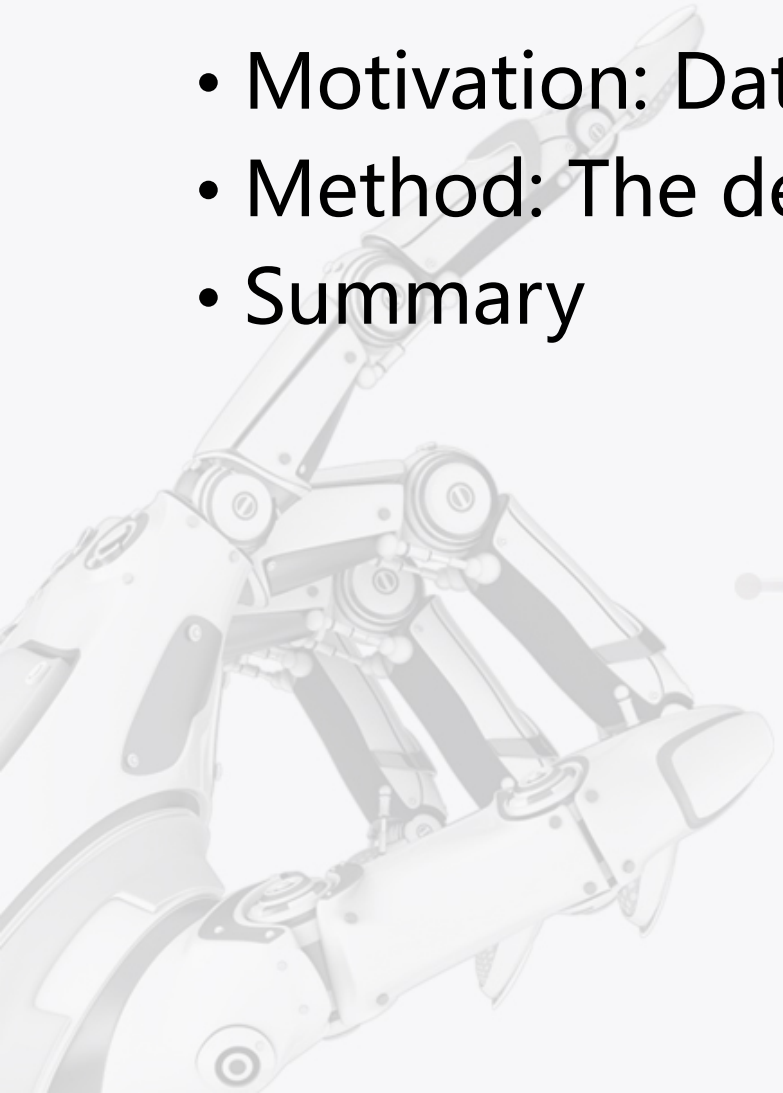
- Working in Baidu
- Focus on AIOps
  - Datacenter network monitoring
  - Time-series anomaly detection
  - Service diagnosis



BaiduIOP

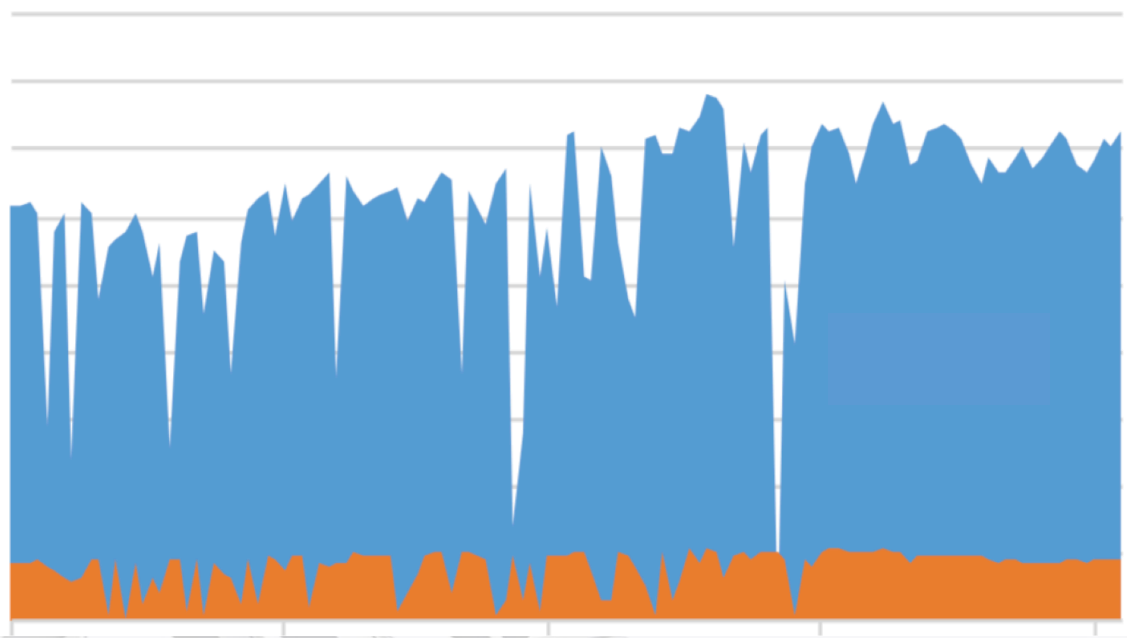
Outline

- Motivation: Datacenter network monitoring is important
- Method: The design of NetRadar
- Summary

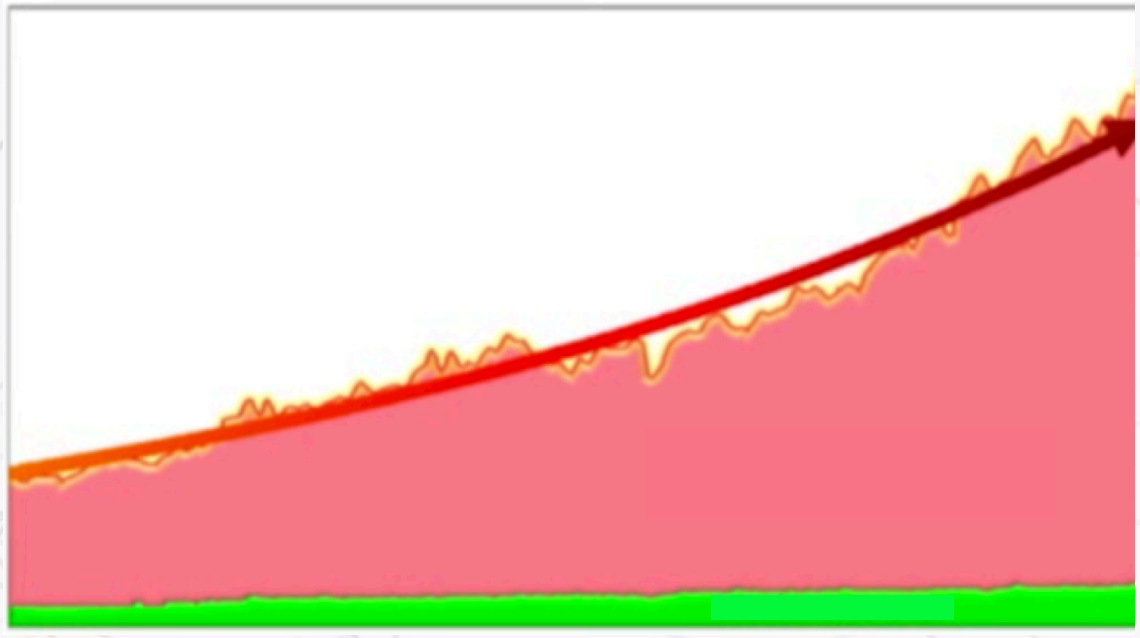


BaiduIOP

The internal traffic is massive

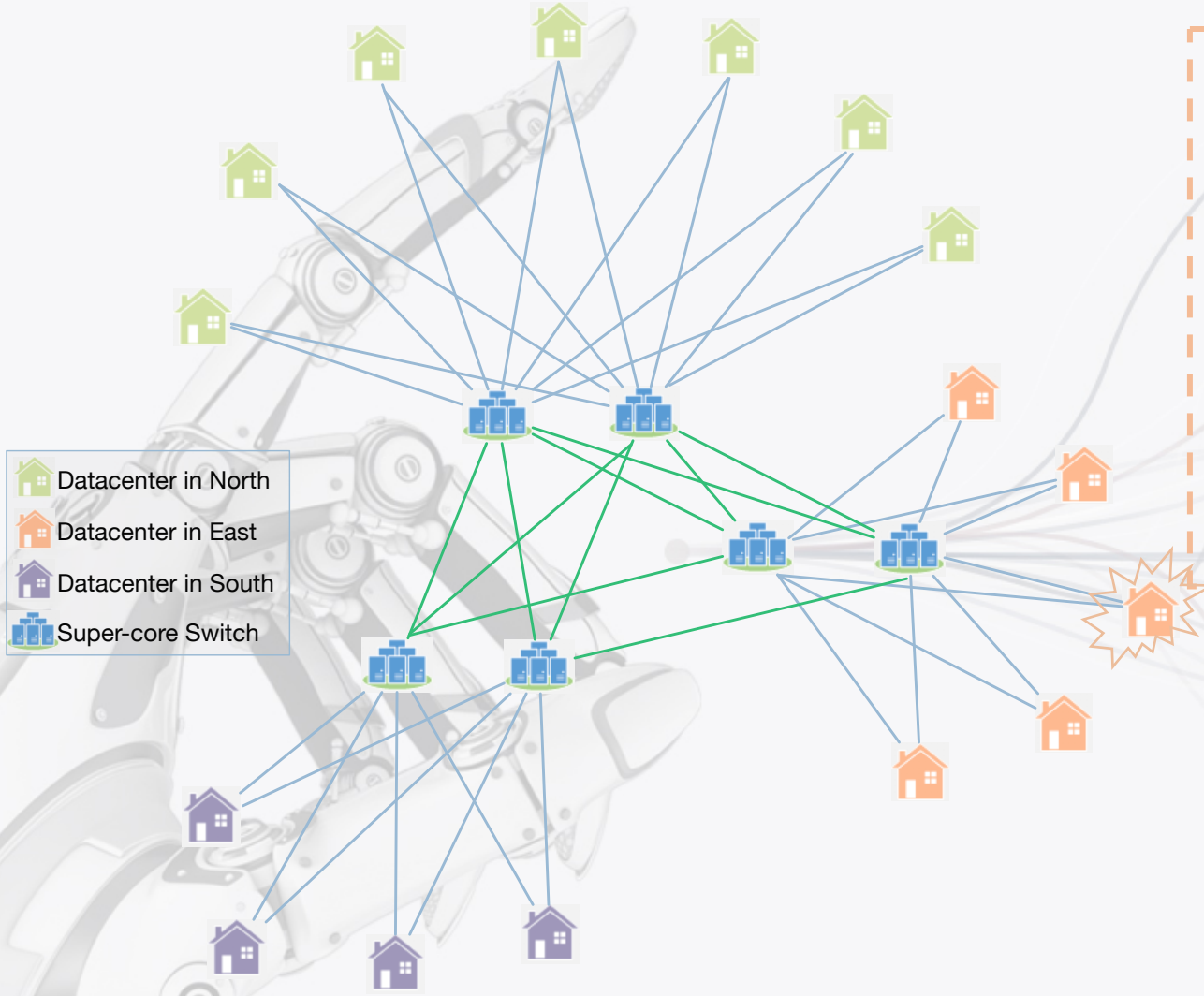


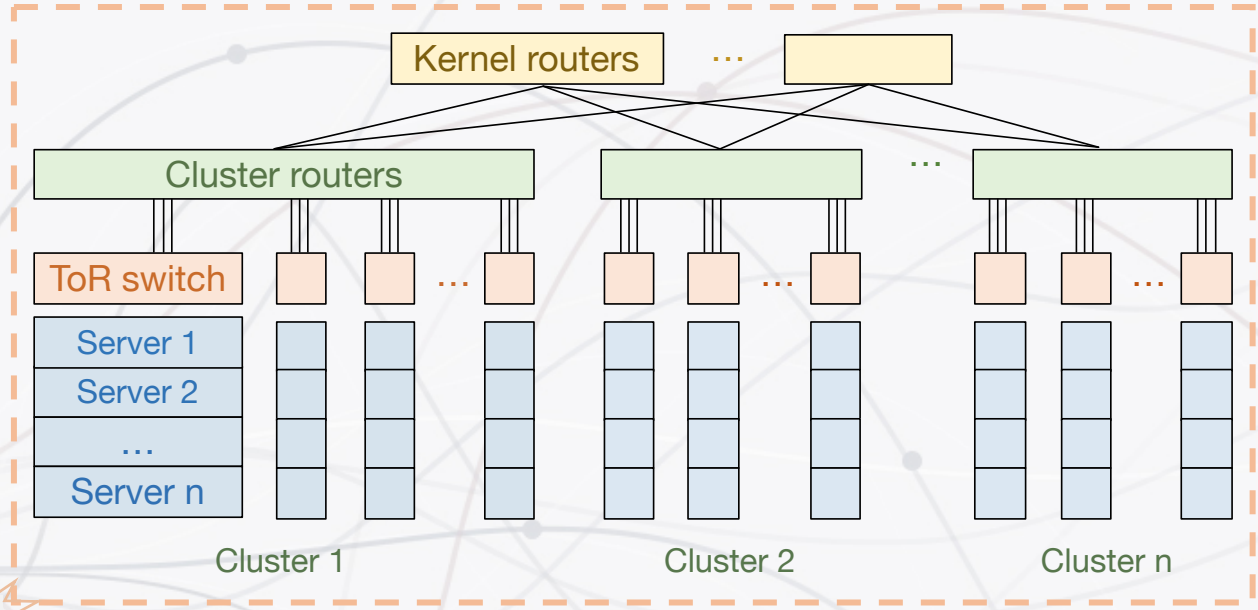
Traffic distribution in Baidu



Traffic distribution in Facebook

Baidu network architecture is complex



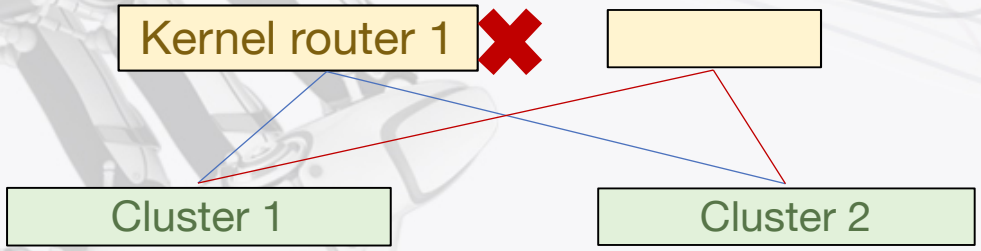


- More than 30 datacenters
- Tens of thousands of network devices
- Hundreds of thousands of servers

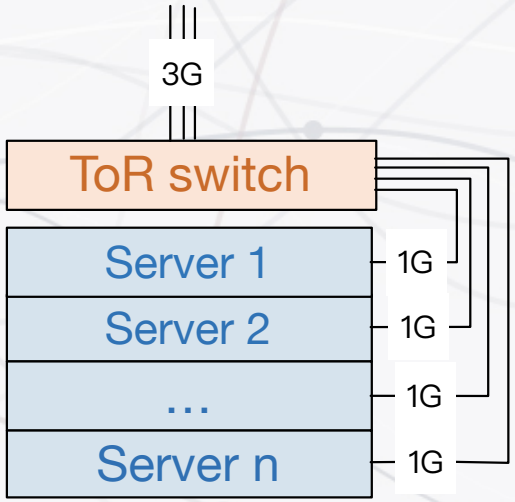
Baidu IOP

Network failure is common

- Device failure happens most of time
  - Device's availability is 99.99% (52 min a year)
  - Probability of failure-free : $0.9999^{10,000} \approx 0.37$
- Redundancy can't avoid network failure entirely
  - Route convergence takes time
  - Congestion due to oversubscription



Network failure
before convergence completed

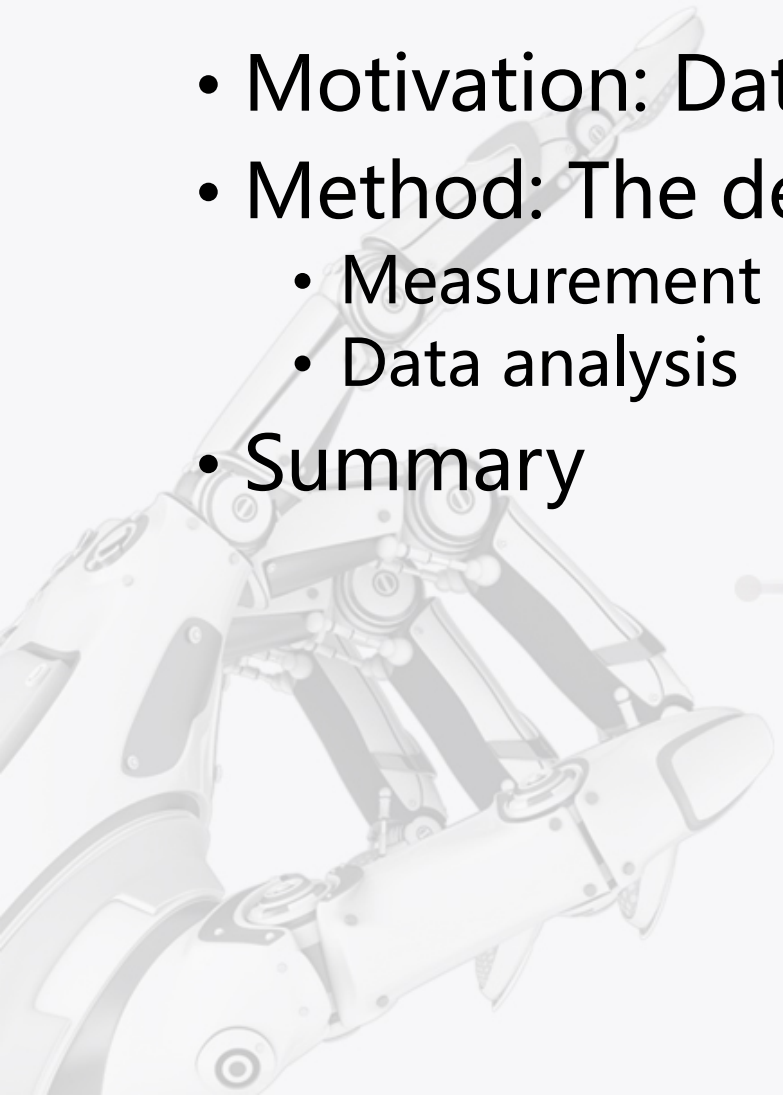


Network congestion
when traffic is more than 3G

Baidu IOP

Outline

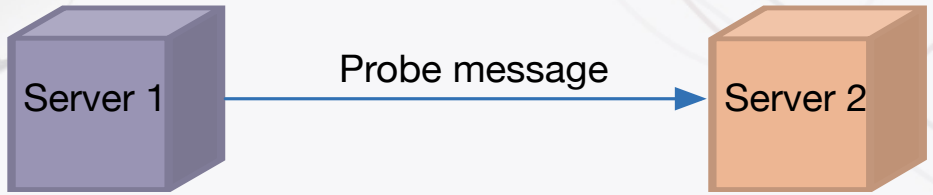
- Motivation: Datacenter network monitoring is important
- Method: The design of NetRadar
  - Measurement
  - Data analysis
- Summary



BaiduIOP

End-to-end monitoring is significant to SRE

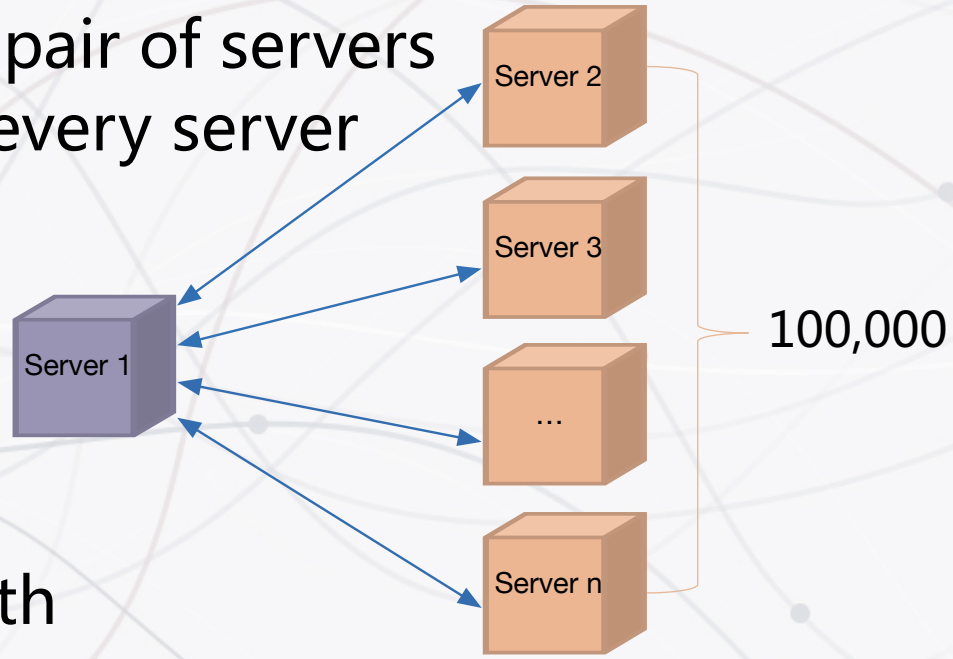
- SRE cares about network issues that applications perceived
- Device monitoring helps little to SRE
  - Device availability != application availability
    - Device fails but applications work well: redundancy
    - Application perceives network issue but devices are healthy: congestion
      - **36%** issues are caused by congestion
- End-to-end probing meets SRE's need
  - Perceives the same network issues with applications



BaiduIOP

Sampling is necessary in the Baidu scenario

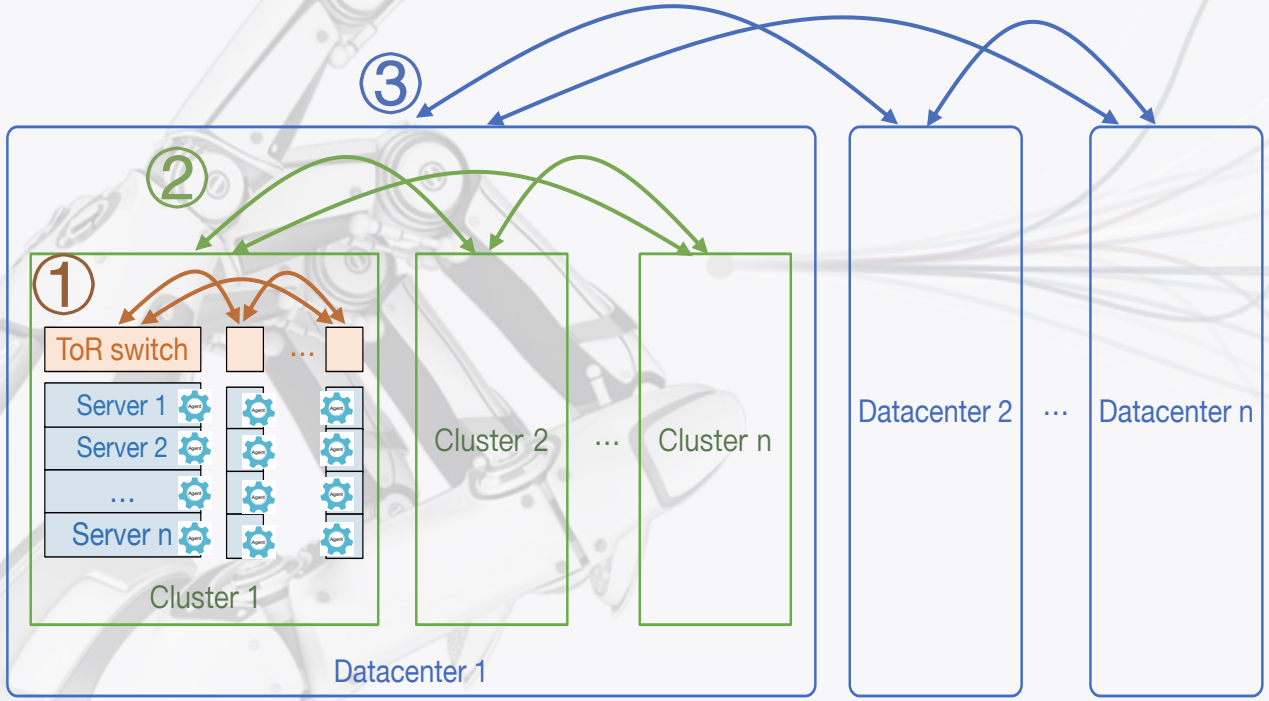
- Simple way costs too much
  - Method: probe between every possible pair of servers
  - Network cost: 533Mbps bandwidth for every server
    - Hundreds of thousands of servers
    - Probing every 30 seconds
    - Probing message size: 10KB
    - Cost: 100,000 * 2 * 10KB / 30s = 533Mbps
- Requirement of sampling
  - Sufficient samples for every network path



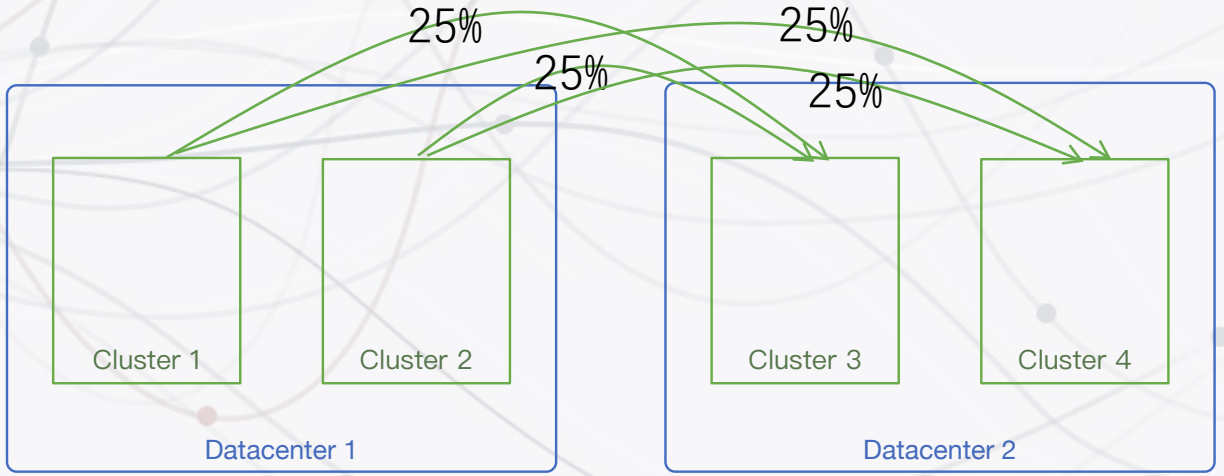
BaiduIOP

Hierarchical and evenly distributed sampling to ensure sufficiency

- Hierarchical probing
  - Consider network architecture



- Samples evenly distributed across the paths
  - Assign similar amount samples for each path



Data analysis is needed to close the gap between observations and SRE's concern

- End-to-end probing results
  - Reachability between server pairs

Level	SrcUnit	DstUnit	SrcIp	DstIp	Reachable
ToR	ToR 1	ToR 2	SIP_1	DIP_1	Success
Cluster	Cluster 1	Cluster 2	SIP_2	DIP_2	Fail
…	…	…	…	…	…
Datacenter	Datacenter 1	Datacenter 2	SIP_n	DIP_n	Success

- SRE's concern
  - ToR or Cluster or Datacenter failure?
  - Failure between Clusters or Datacenters?

BaiduIOP

Simple idea: detection for every possible failure SRE concerned

- Enumerate all failures for given architecture
- Detect for every possible failure separately
  - An example：detection for Cluster 1
    - Gathering all probing result
      - SrcUnit = Cluster 1 or DstUnit = Cluster 1
    - Count probing times(n) and drop times(k)
    - Is k too large for given n?
      - Errors detection
      - Alert condition: $P(X \geq k)$ is very small

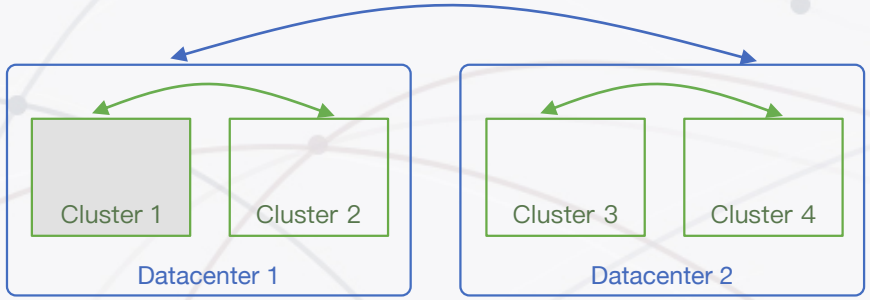
$$P(X = k) = C_n^k (1 - p_0)^{n-k} p_0^k$$
$$P(X \geq k) = \sum_{i=k}^{n} P(X = i)$$

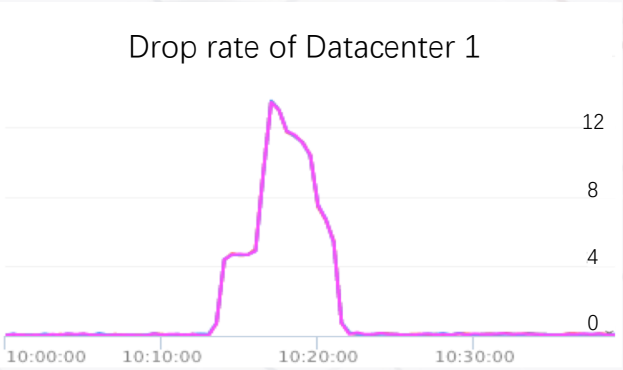
SrcUnit	DstUnit	SrcIp	DstIp	Reachable
ToR 1	ToR 2	SIP_1	DIP_1	Success
Cluster 1	Cluster 2	SIP_2	DIP_2	Fail
ToR 2	ToR 3	SIP_3	DIP_3	Success
Cluster 2	Cluster 1	SIP_4	DIP_4	Success
Cluster 3	Cluster 1	SIP_5	DIP_5	Fail
…	…	…	…	…

BaiduIOP

Detection separately triggers false positive

- Issue introduction
  - Cluster 1 failure with 25% drop rate
- Detect Cluster 1 failure successfully
  - Drop rate of Cluster 1 increased sharply, Alert
- False positive: Datacenter 1 failure
  - Half of samples in Datacenter 1 are from Cluster 1
  - Drop rate of Datacenter 1 increased because of Cluster 1 failure





Multi-dimensional data analysis helps to eliminate false positives (1)

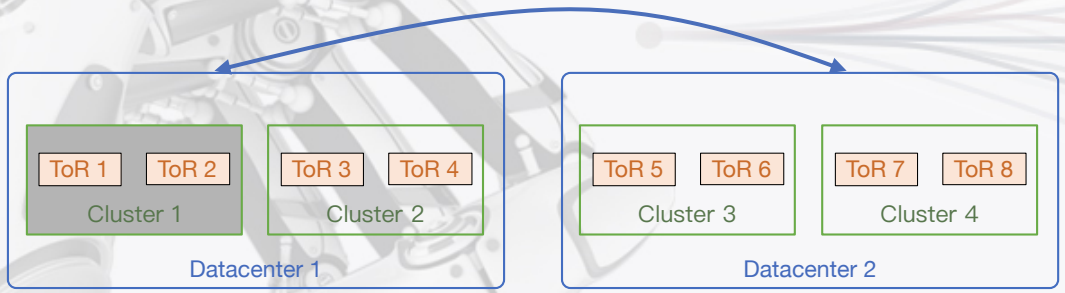
- Label probing result with tags
  - Tags name: SrcIp、SrcToR、SrcCluster、SrcDatacenter、DstIp、DstToR、DstCluster、DstDatacenter
  - Tags value: depend on SrcIp、DstIp
- Multi-dimension data
  - Dimension: tag name
  - Dimension value: tag value

Dimension combination	Reachable
SrcIp=SIP_1&SrcToR=ToR_1&SrcCluster=$\mathrm{Cluster}_1$&SrcDatacenter=$\mathrm{Datacenter}_1$ &DstIp=DIP_1&DstToR=ToR_3&DstCluster=$\mathrm{Cluster}_2$&DstDatacenter=$\mathrm{Datacenter}_1$	Success

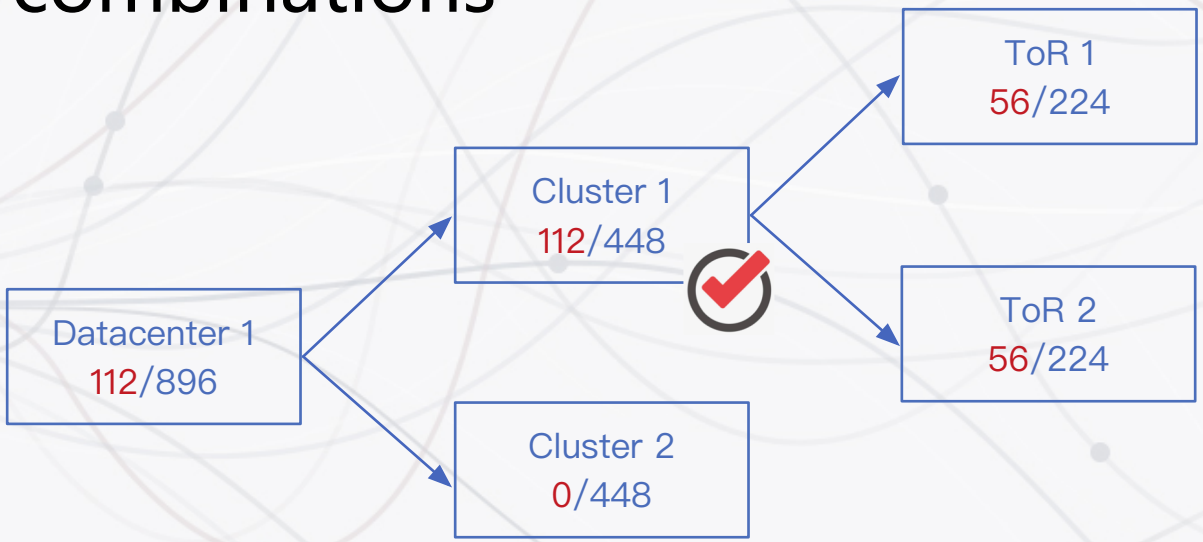
BaiduIOP

Multi-dimensional data analysis helps to eliminate false positives (2)

- Failure can be represented by dimension combination
  - SrcCluster=$\mathrm{Cluster}_1$&DstCluster=$\mathrm{Cluster}_2$
- Feature of failure dimension combinations
  - High contribution
  - High uniformity



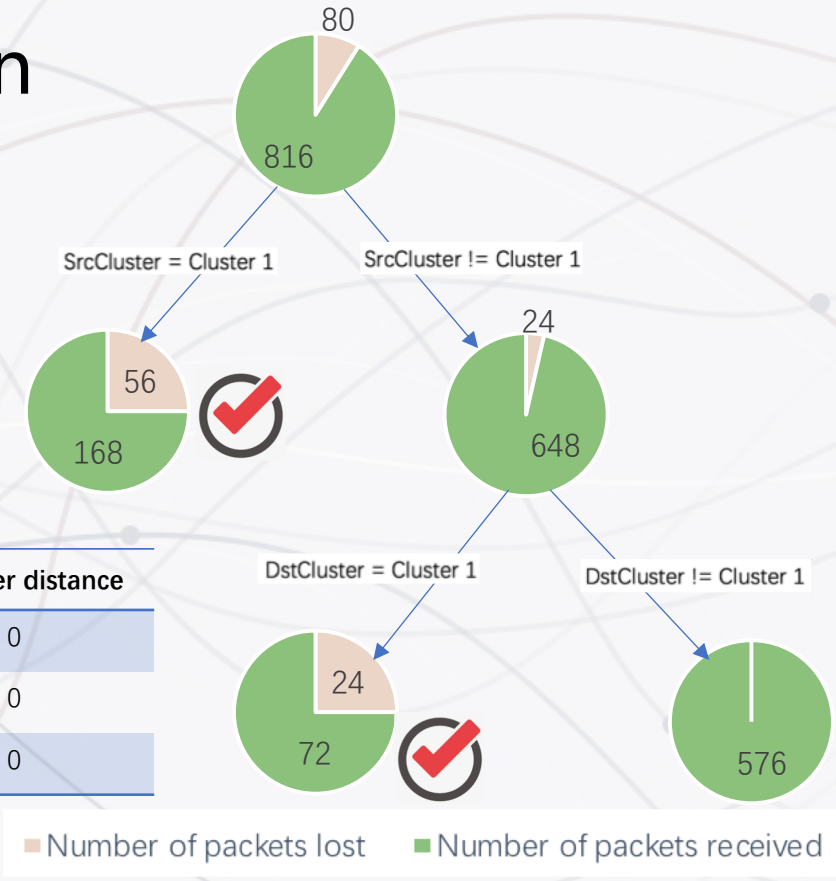
Cluster 1 failure with 25% drop rate



An efficient multi-dimensional data analysis algorithm based on decision tree

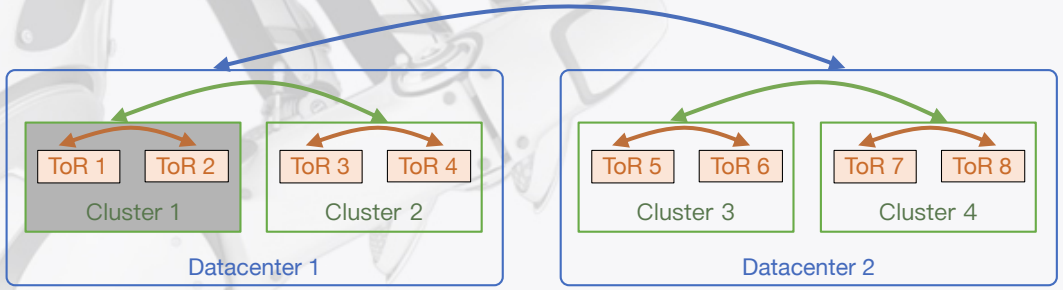
- Select the most distinguishing dimension
  - Hellinger distance

Dimension combinations	Hellinger distance
SrcCluster=Cluster 1	0.363
SrcDatacenter=Datacenter 1	0.239
SrcToR=ToR 1	0.216
SrcToR=ToR 3	0.030
...	...



- Stop condition for splitting
  - Uniformly distributed
  - Little drops

Dimension combinations	Hellinger distance
SrcToR=ToR 1	0
SrcToR=ToR 2	0
...	0



Baidu IOP

Summary

- Measurement
  - End-to-end probing is used in consideration of SRE's need
  - Sampling is necessary in the Baidu scenario
    - All possible paths are evenly distributed
- Data analysis
  - Detection for every failure separately may trigger false positive
  - Multi-dimensional data analysis can eliminate false positive
    - Basic idea of multi-dimensional data analysis
    - An efficient multi-dimensional data analysis algorithm

Thanks

