Identifying Anomalous Traffic Using Delta Traffic

Tsuyoshi KONDOH and Keisuke ISHIBASHI
Information Sharing Platform Labs.
NTT

Flocon2008, January 7–10, 2008, Savannah GA

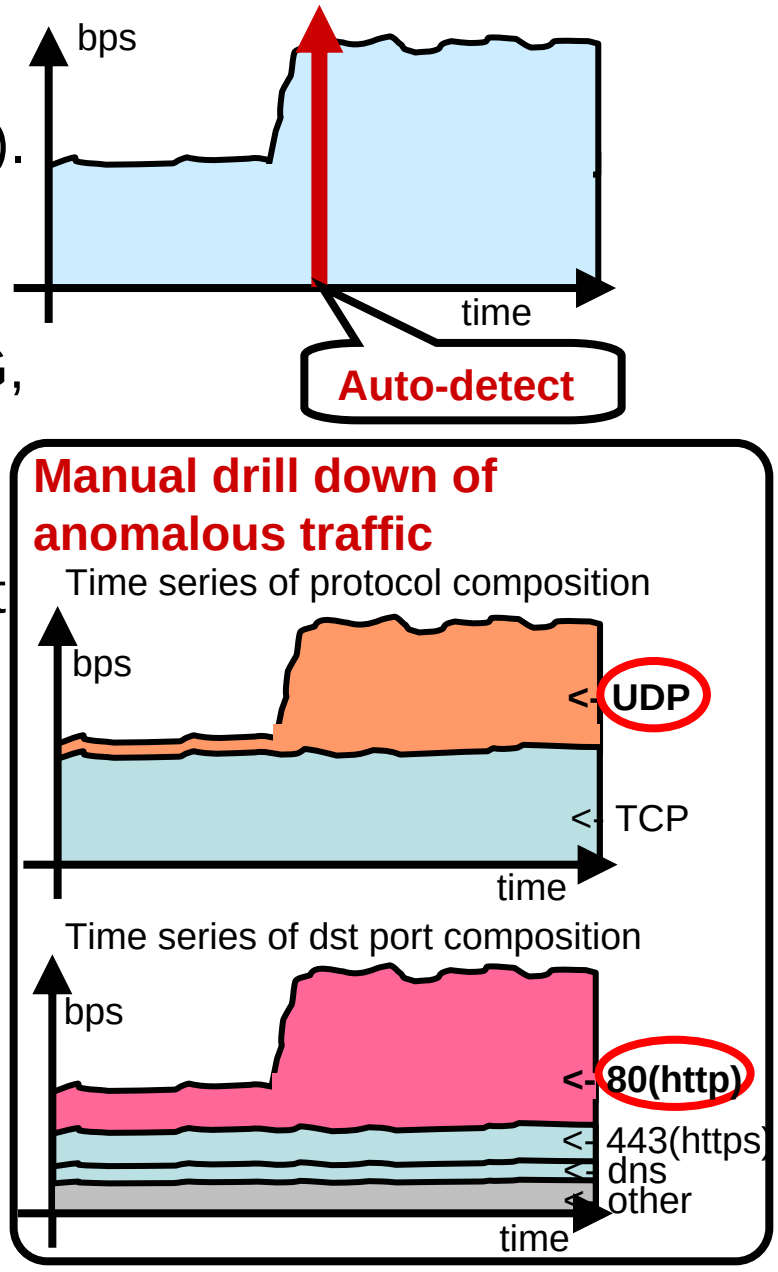
Outline

- Background and Motivation
  - Identifying anomalous traffic is the missing piece.
- Our Technique: DELTAA
  - Concepts
    1. Extract anomalous traffic as the delta of normal and anomalous time periods.
    2. Auto-aggregate extracted anomalous traffic.
  - Operation of our technique
    - Show the step by step operation of our technique.
- Evaluation
  - Evaluation using synthesized DDoS traffic.
- Summary

Background and Motivation

- Monitoring of traffic volumes is widely used for network operation (e.g. MRTG).
- Many techniques for detecting anomalous volume change have been proposed (NBAD, Holt-winters in MRTG, … etc.).
- Some tools to mitigate damage from anomalous traffic. (e.g. drop/rate limit at router, detour to Cisco Guard, etc.)

- However, accurate mitigation needs accurate ACLs (ACL set).
- But, Generating accurate ACL set requires manual drill down by operator.
  - **It's too costly.**



Our Technique: DELTAA

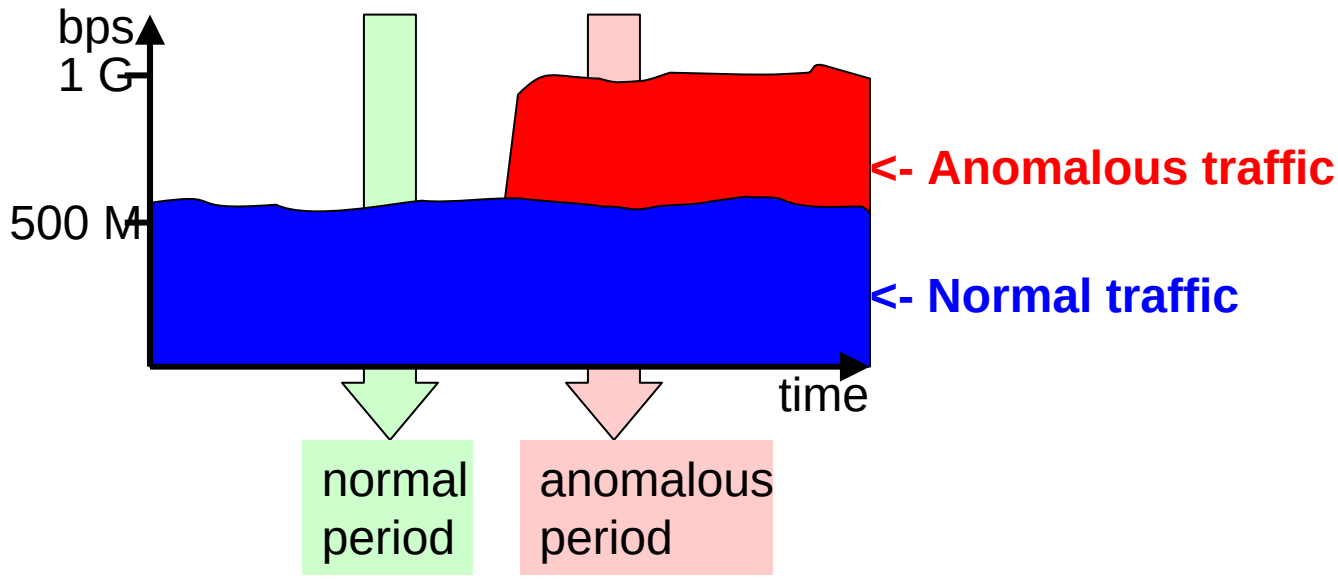
- **DELTAA** outputs ACL set for filtering or rate limiting to mitigate the damage from anomalous traffic.
  - DELTAA: Delta Traffic Automatic Aggregator

- Three concepts of DELTAA:
  1. Reveal anomalous traffic using delta traffic, between normal and anomalous periods.
  2. Aggregate delta traffic and generate optimized ACL set on single dimensions.
     - Dimension means source IP address, destination IP address, protocol or port numbers.
  3. Generate multi-dimensional ACL set by integrating each single dimensional ACL set.

Concept #1:
(1) Definition of "Normal" and "Anomalous" Traffic

Throughout this presentation, I use the following definitions.

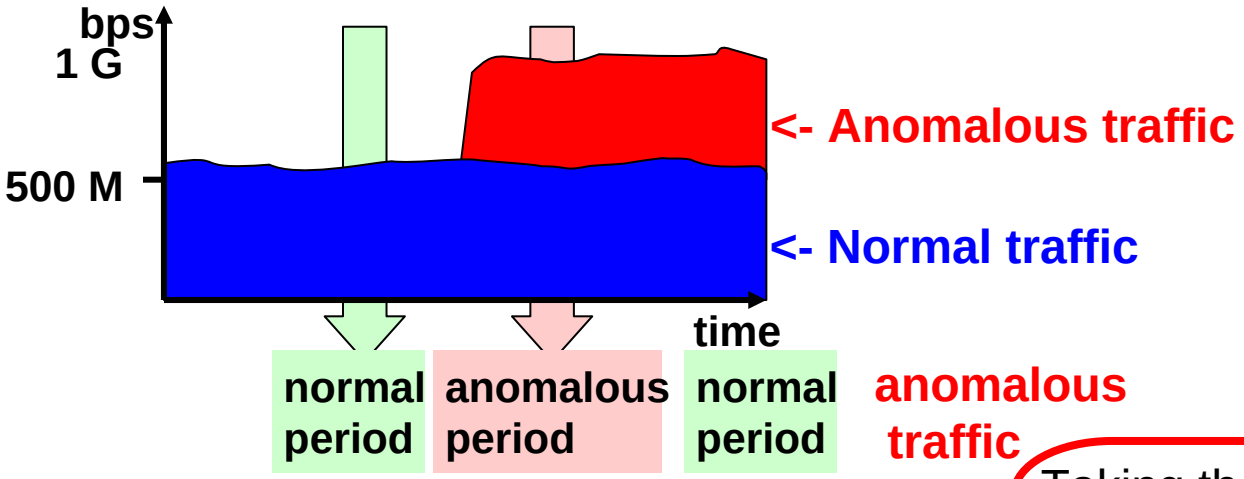
1. Anomalous traffic: Traffic that causes a change in traffic volume (bps/pps/fps).
   - BitTorrent and server intrusion are out of scope because they always exist or do not cause a change in traffic volume.
2. Normal period: Period when traffic volume is normal.
3. Anomalous period: Period when traffic volume is anomalous.



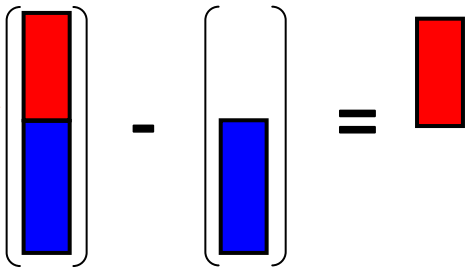
Concept #1 :
(2) Reveal Anomalous Traffic

- Make two assumptions
  1. traffic of normal period = normal traffic
  2. traffic of anomalous period = normal traffic + anomalous traffic

- anomalous traffic = traffic of anomalous period – traffic of normal period



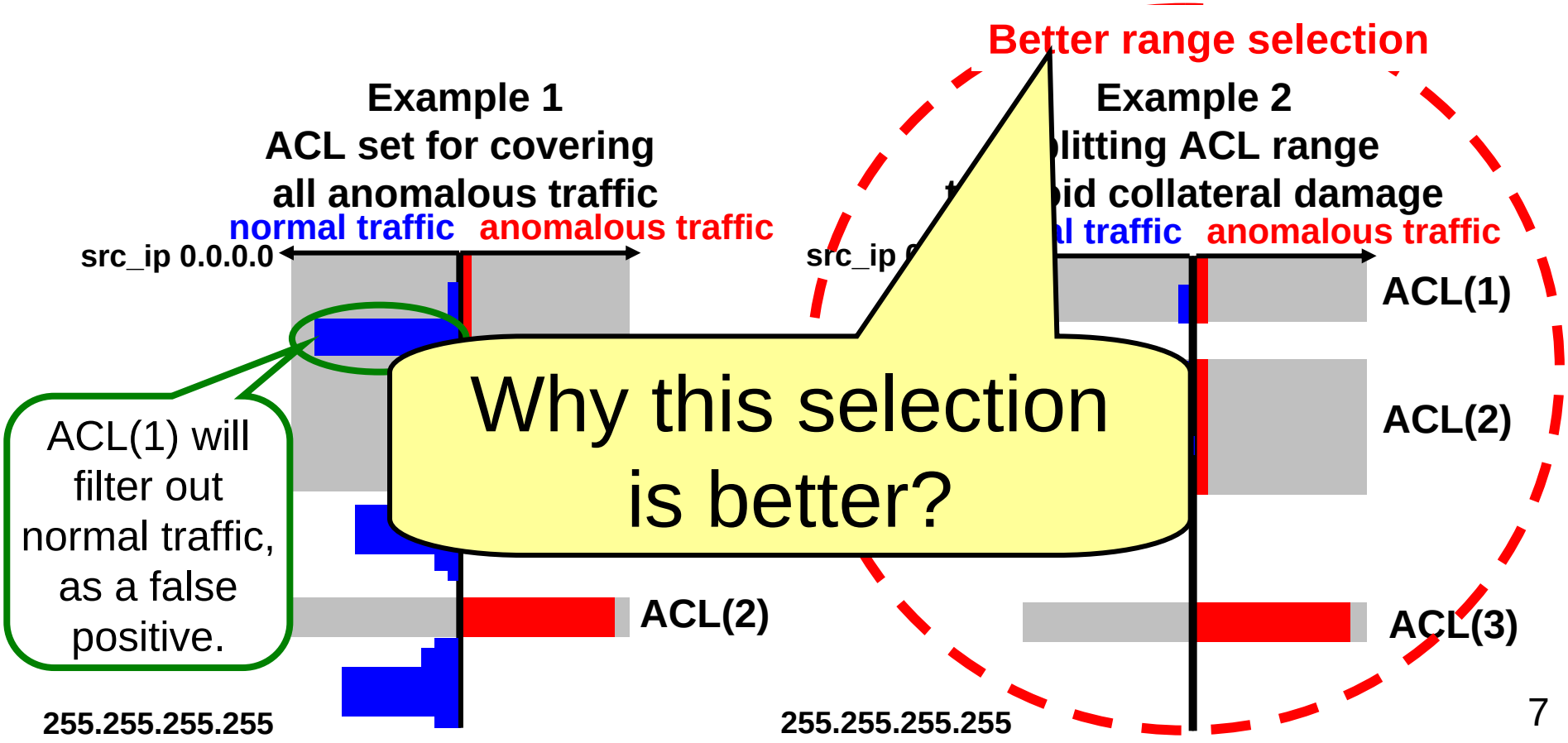
Extracting anomalous traffic from “traffic of anomalous period” is difficult because it is a mixture of normal and anomalous traffic.



Taking the delta between “traffic of normal period” and that of anomalous period, we can effectively extract anomalous traffic.

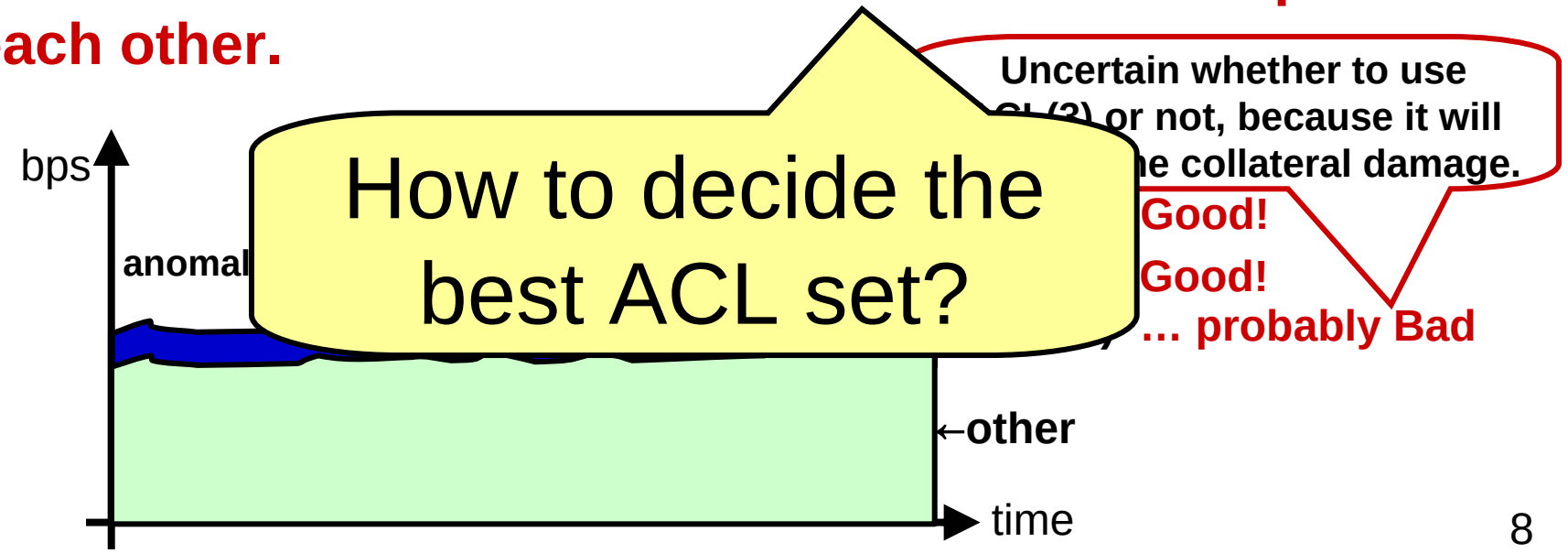
Concept #2:
Auto-aggregate Delta Traffic

- Our technique expresses anomalous traffic with some number of ranges.
  - For example source IP address ranges.
  - The ranges should be optimal for filtering.



Criteria of "Goodness"

- We introduce three criteria of identification.
  1. Coverage ratio: (1 – FNR)
     Maximize filtered anomalous traffic
  2. Collateral (damage) ratio: (FPR)
     Minimize filtered (normal) legitimate traffic
  3. **Number of ACLs**:
     ACL entry budget is limited, so fewer ACLs is better.
- **These three criteria have a trade-off relationship with each other.**



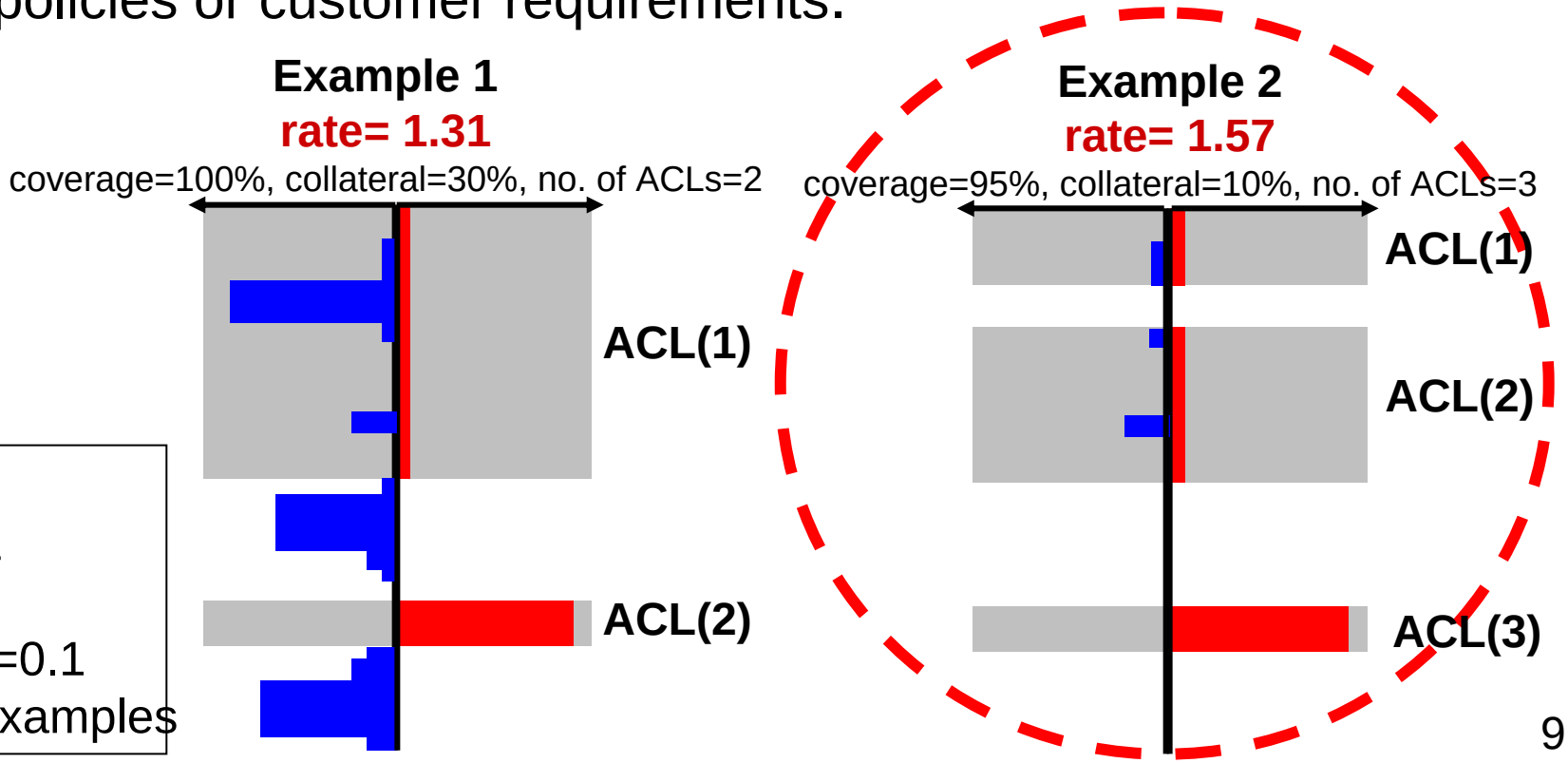
Evaluation Formula for Goodness

To decide the best ACL set, we introduce this formula:

$$\text{coverage}: cov,\ \text{collateral ratio}: coll,\ \text{no. of ACLs}: n$$

$$\text{rate} = \frac{(\beta - \alpha) + \alpha \cdot cov - \beta \cdot coll}{n^{\gamma}} \quad (\alpha, \beta, \gamma : \text{weighting coefficients})$$

– By tuning the weighting coefficients, we can reflect network policies or customer requirements.



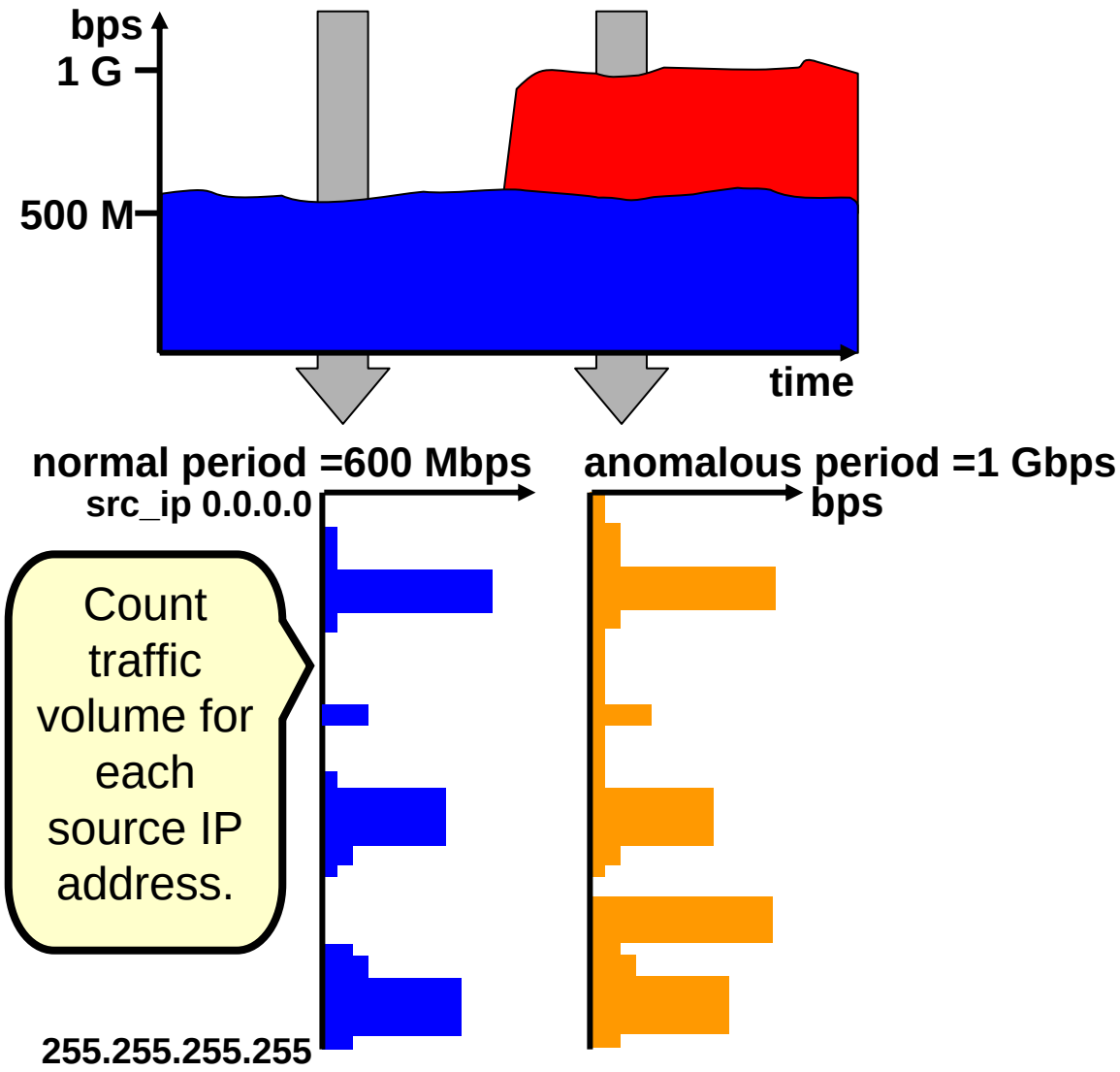
Use
-alpha=1
-beta=2
-gamma=0.1
for this examples

Step by Step Explanation of Our Technique

- Following seven pages show the step by step operation of our technique including above two concepts and concept #3.

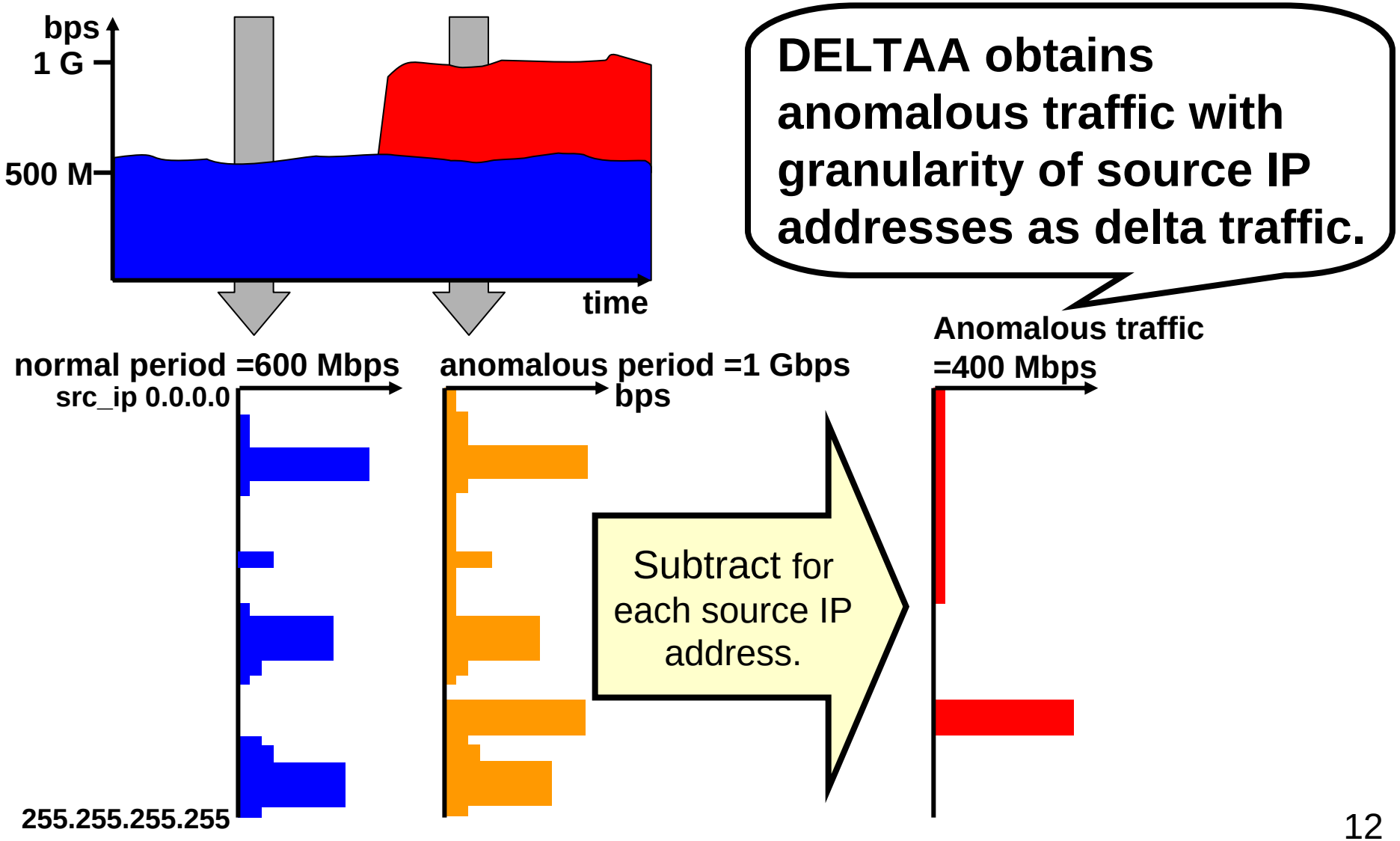
Step 1: (1) Counting Up

Count traffic of both normal and anomalous periods for each source IP address.



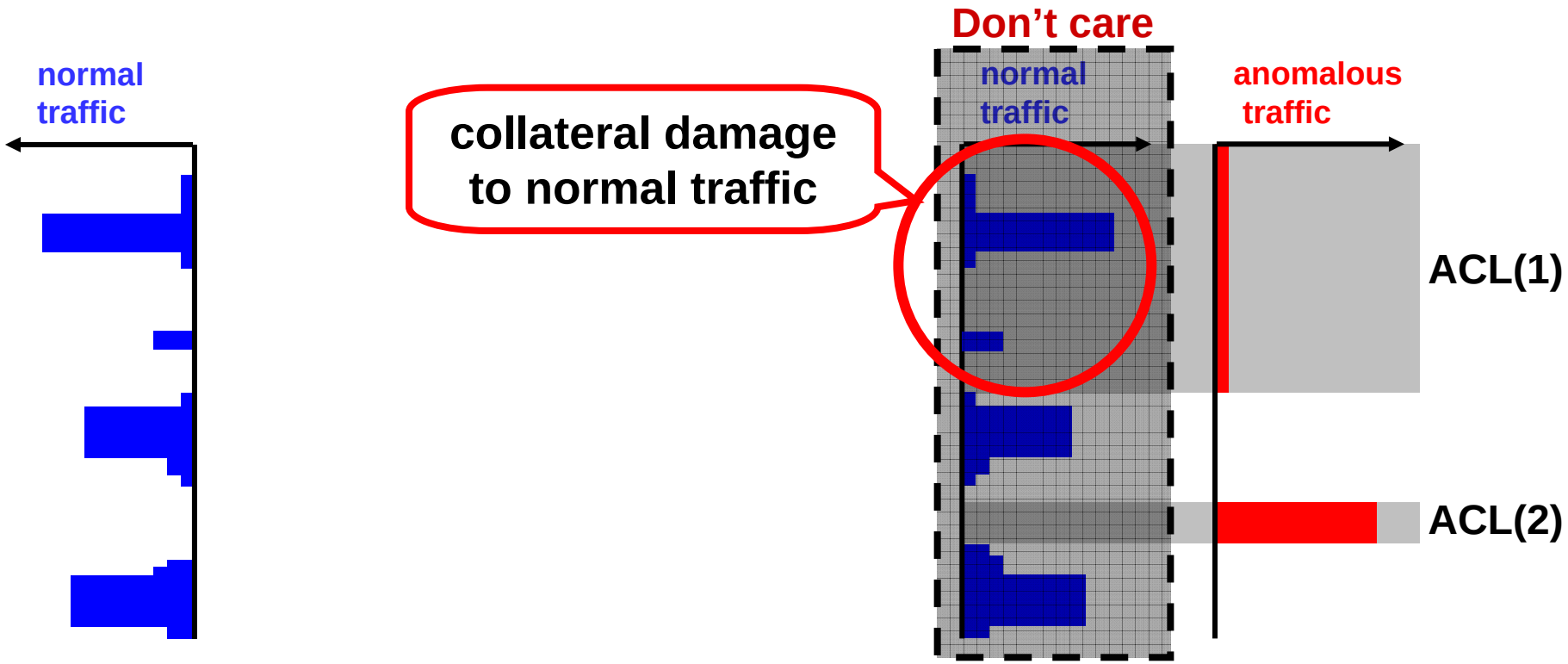
Step 1: (2) Making Delta Traffic

Make delta traffic by subtracting traffic of normal period from that of anomalous period.



Step 2: (1) Deciding ACL Set as IP Address Ranges

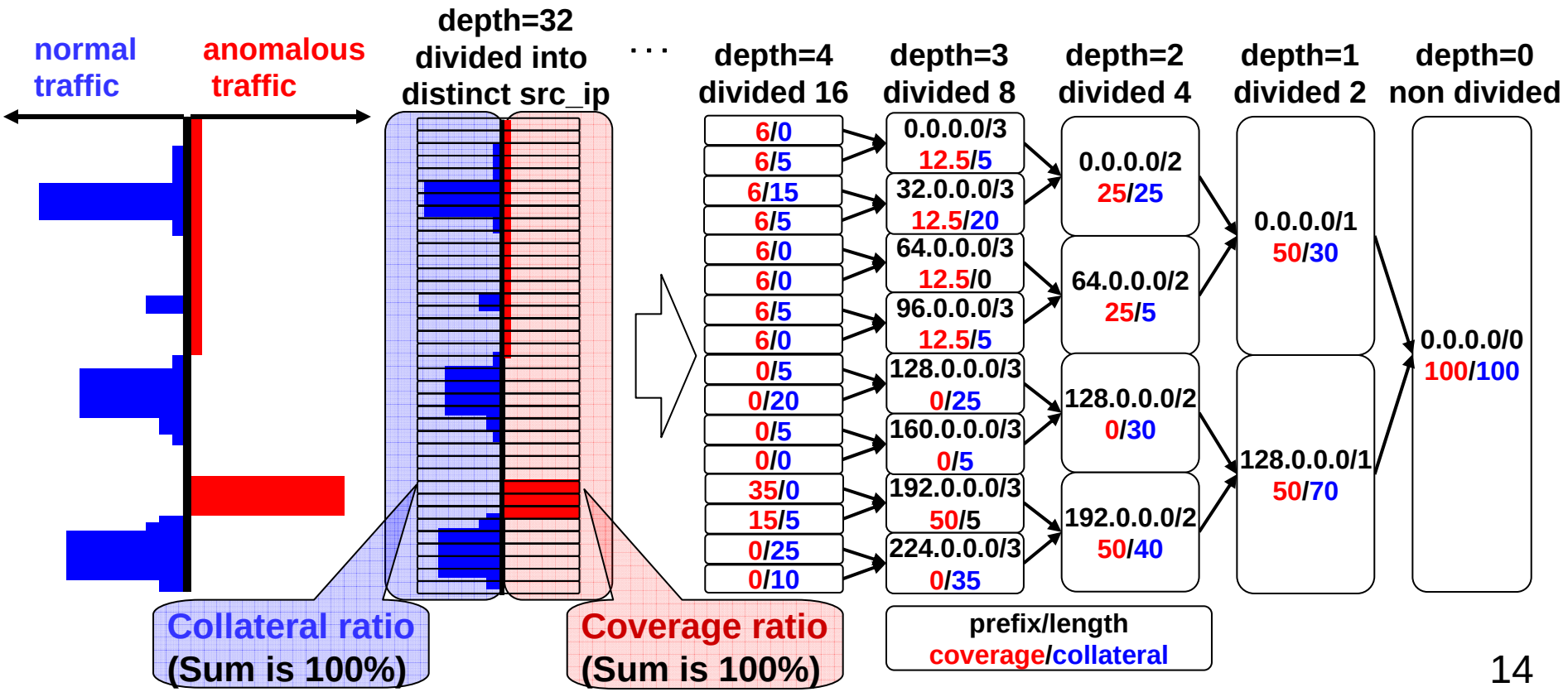
- When using anomalous traffic information only, collateral damage cannot be avoided.
  - Causes miss-filtering of normal traffic.
- So, we need to use information on both normal and anomalous traffic.



Step 2: (2) Building Tree of Normal and Anomalous Traffic

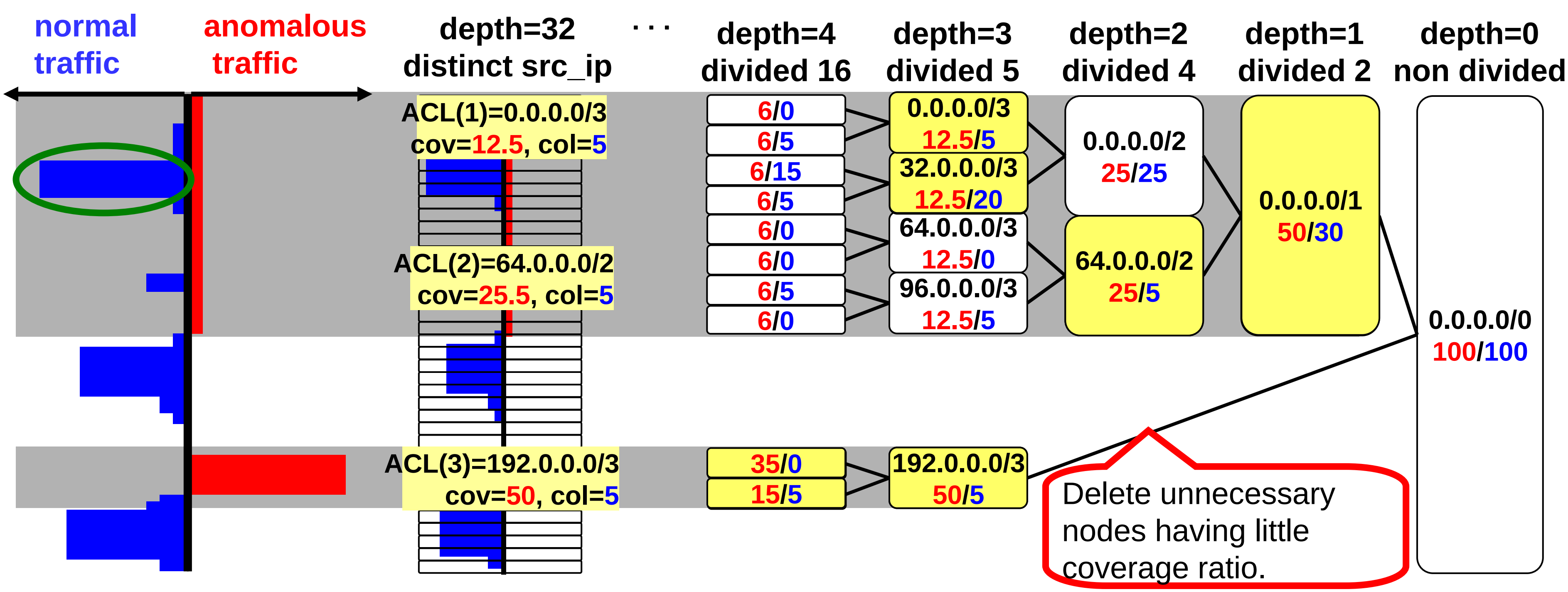
Making Traffic Tree

- Build up from individual source IP addresses (depth=32).
- Each node has information about coverage and collateral ratio.
  - Collateral ratio: normal traffic of the node ÷ total normal traffic
  - Coverage ratio: anomalous traffic of the node ÷ total anomalous traffic
- Make parent nodes by merging child node information.



Step 3: Selecting Best Node Set from Traffic Tree

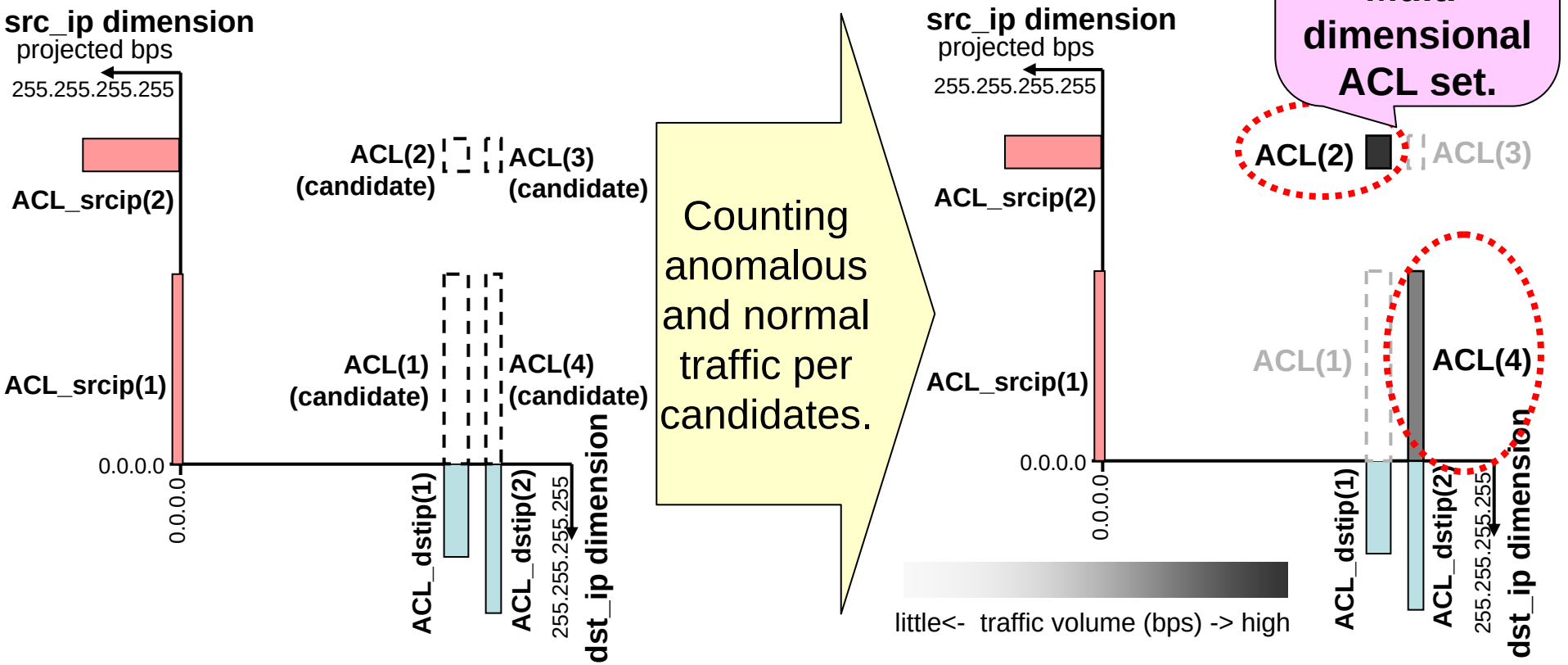
1. To reduce search space, delete unnecessary nodes.
   - Unnecessary node: node which having little coverage ratio (little anomalous traffic) or little difference from its descendant nodes.
2. Search best node combination by applying the formula for all non-overlap node combinations in a brute force way.
   - Best node combination = **Best ACL set for source IP dimension**



Example (overlapping slide animation text): **rate= ... : Can cover ... anomalous traffic, but ... collateral ... with little collateral ... must be aggregated to /3 (depth=3)**

Concept #3 Generate Multi-Dimensional ACL Set

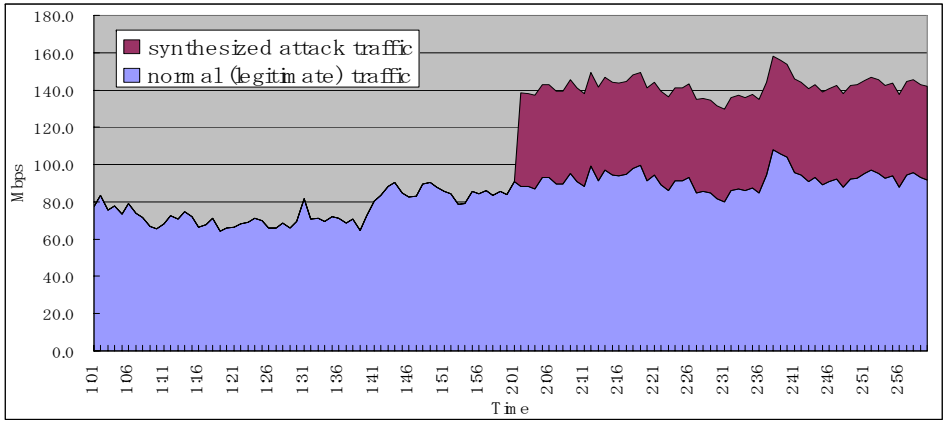
- Generate single dimensional ACL set in parallel.
  - ‘source IP’, ‘destination IP’ , ‘protocol’, ‘source port’ and ‘destination port’
- Make candidates of multi-dimensional ACL sets as a product sets of each dimension.
- Count anomalous/normal traffic for every candidates.
- Select best combination of candidates in terms of goodness score.



Evaluation and Results: Test Data Set

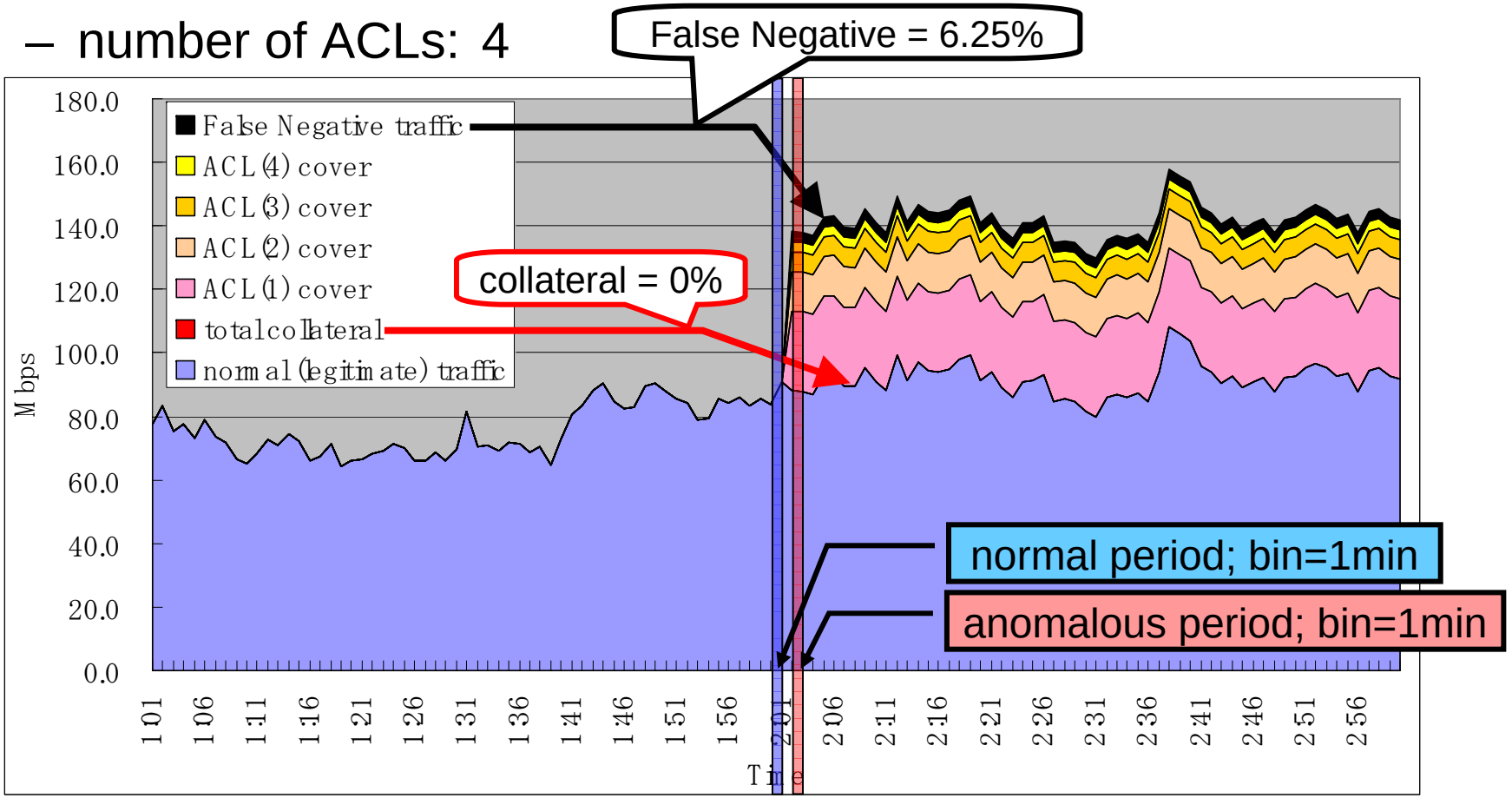
- Normal traffic: publicly available traffic data, captured on transpacific line (100 Mbps)
- Anomalous traffic: injected synthesized DDoS attack traffic
  - Mimic large DDoS attack
    - We choose source/destination IP addresses that have large normal traffic, because simple identification would cause collateral.
    - Destination: Popular server appeared in normal traffic
    - Source: Choose IP address block (/16) from which volume of normal traffic to the destination is largest.
- Test how well our technique can extract the injected anomalous traffic.

Use the "weighting coefficients"
-alpha=1 (weight for coverage)
-beta=10,000 (weight for collateral)
-gamma=0.0001 (weight for no. of ACLs)
to avoid collateral damage



Evaluation and Results: Results (1)

- Results: We get four ACLs (Four ACLs are one set.)
  - coverage: 93.75%
  - collateral: 0.00%
  - number of ACLs: 4



Time series of traffic with output ACLs displayed in separate colors

Evaluation and Results: Results (2) OUTPUT

basetime_len= 60.0 (sec) : (1168362060.0 - 1168362120.0) **basic information**
anomtime_len= 60.0 (sec) : (1168362180.0 - 1168362240.0)
base_total_bps= 89,121,539.5
anom_total_bps= 137,729,812.7
diff_total_bps= 48,608,273.2 +54.5 %

1-D_OUTPUT: PROTOCOL= 6 coverage= 100.42 collateral= 95.52 **single dimensional**
1-D_OUTPUT: SRC_PORT= high coverage= 108.27 collateral= 33.42 **identification**
1-D_OUTPUT: DST_PORT= high coverage= 100.09 collateral= 96.40 **results**
1-D_OUTPUT: **SRC_IP** coverage= 96.43 collateral= 0.00
119.170.0.0/17 coverage= 51.43 collateral= 0.00
119.170.128.0/18 coverage= 25.72 collateral= 0.00
119.170.192.0/19 coverage= 12.86 collateral= 0.00
119.170.240.0/20 coverage= 6.43 collateral= 0.00
1-D_OUTPUT: **DST_IP** coverage= 102.93 collateral= 2.17
134.45.182.70/32 coverage= 102.93 collateral= 2.17

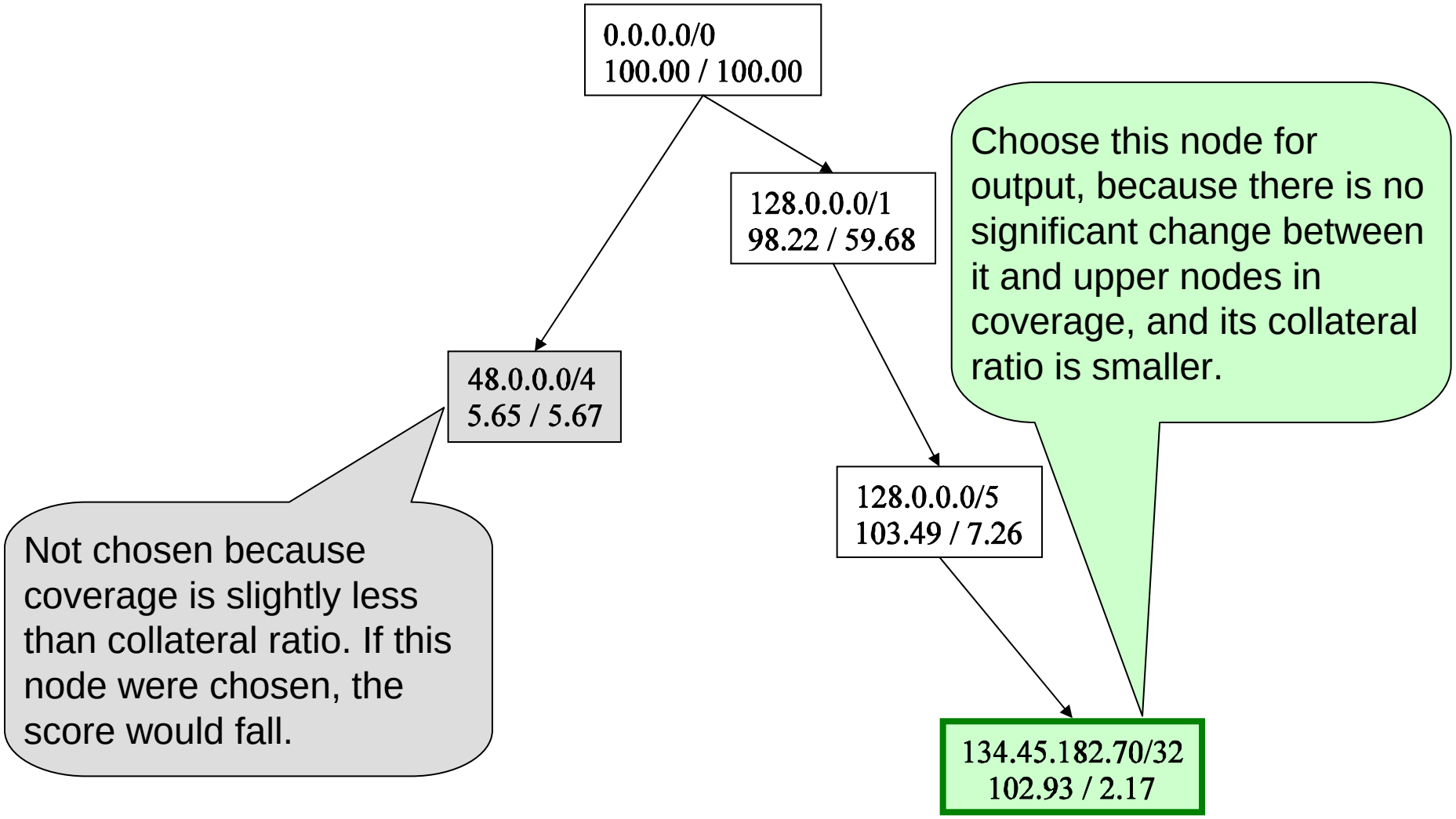
MULTI-DIMENSION_FLOW_OUTPUT coverage= 96.43 collateral= 0.00

flowID_0: cov= 51.43	col= 0.00:	**119.170.0.0/17**	**134.45.182.70/32**	6	high	high
flowID_1: cov= 25.72	col= 0.00:	**119.170.128.0/18**	**134.45.182.70/32**	6	high	high
flowID_2: cov= 12.86	col= 0.00:	**119.170.192.0/19**	**134.45.182.70/32**	6	high	high
flowID_3: cov= 6.43	col= 0.00:	**119.170.240.0/20**	**134.45.182.70/32**	6	high	high

↑ **coverage** ↑ **collateral**: ↑ src_ip ↑ dst_ip protocl scr_port dst_port

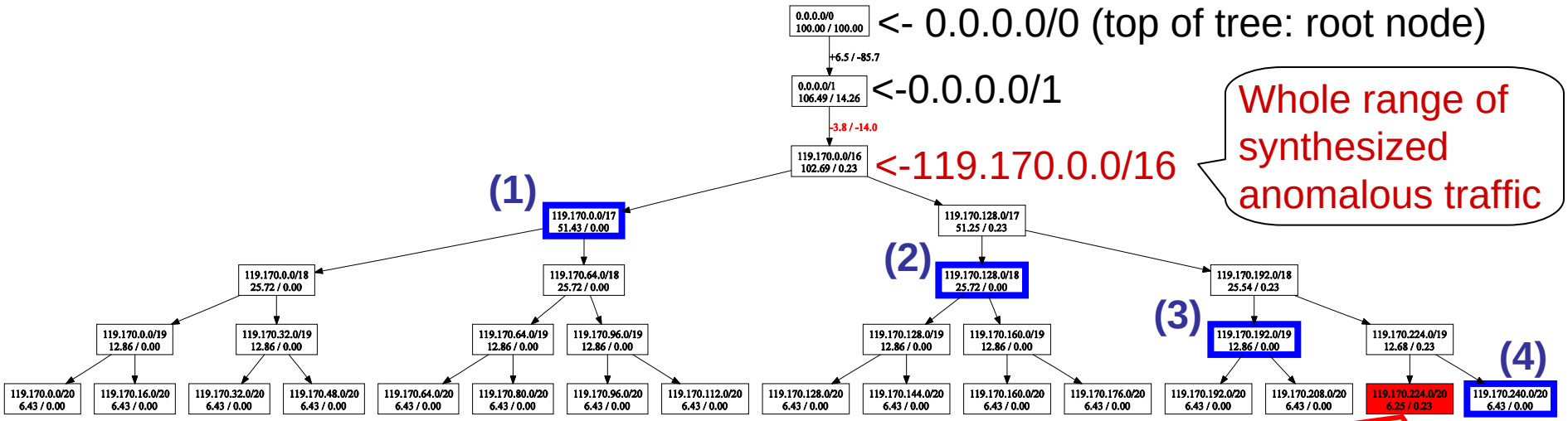
Evaluation and Results (3): Destination IP Tree

1-D_OUTPUT: **DST_IP** coverage= 102.93 collateral= 2.17
134.45.182.70/32 coverage= 102.93 collateral= 2.17



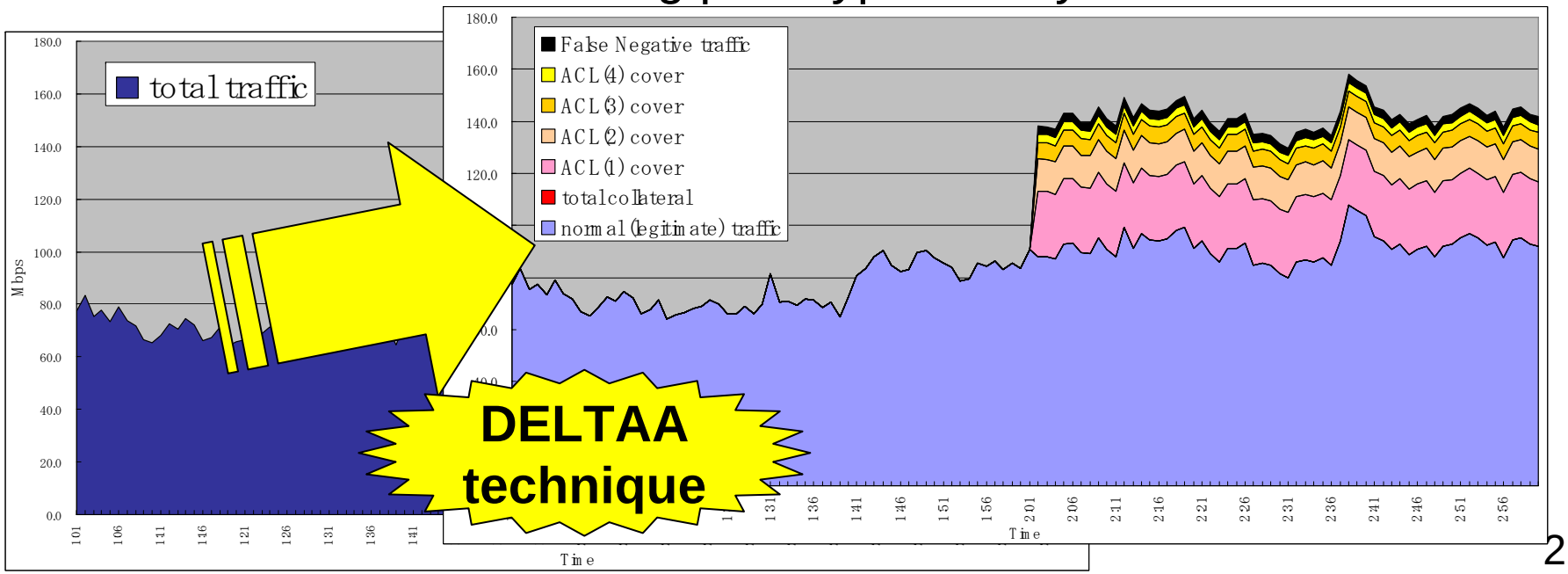
Evaluation and Results (4): Source IP Tree

1-D_OUTPUT: **SRC_IP**		coverage= 96.43	collateral= 0.00
(1)	**119.170.0.0/17**	coverage= 51.43	collateral= 0.00
(2)	**119.170.128.0/18**	coverage= 25.72	collateral= 0.00
(3)	**119.170.192.0/19**	coverage= 12.86	collateral= 0.00
(4)	**119.170.240.0/20**	coverage= 6.43	collateral= 0.00



Summary

- Introduced three criteria of optimal ACL set.
  - for mitigating DDoS attacks on router
- Proposed DELTAA technique: Optimizes trade-off among the these criteria, using normal and anomalous traffic.
- Presented an example of applying DELTAA to extract injected anomalous traffic.
  - Evaluation results using prototype and synthesized data set.



Thank you.

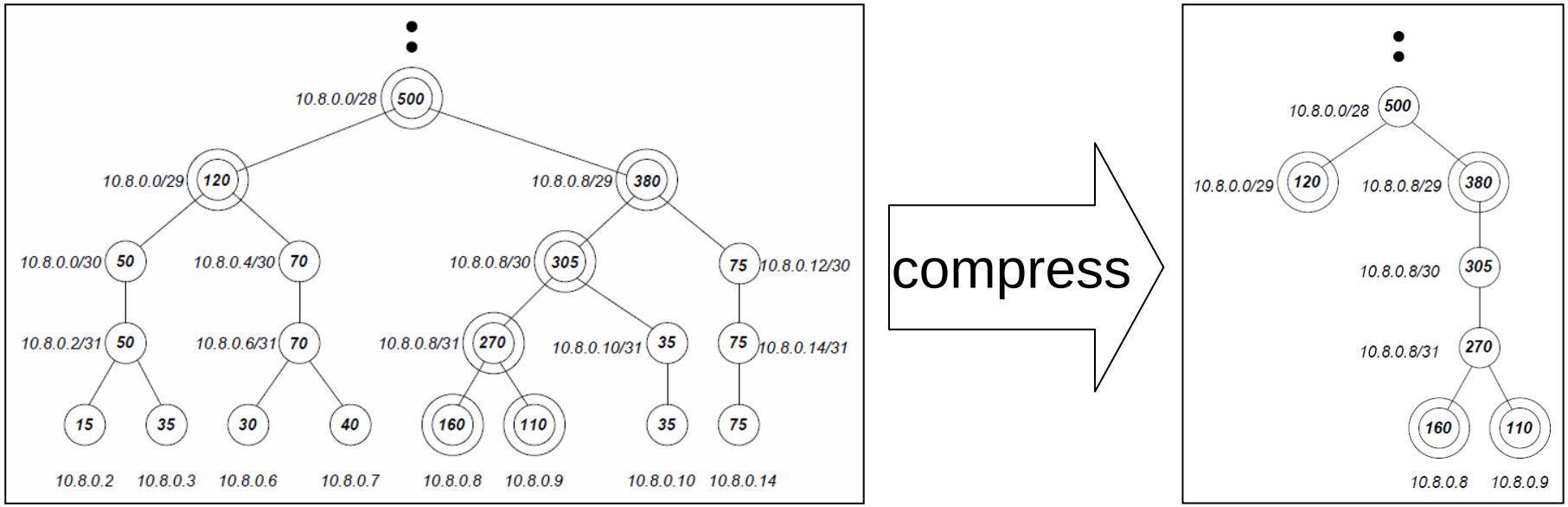
Any questions are welcome.

tsuyoshi.kondoh [at] lab.ntt.co.jp
ishibashi.keisuke [at] lab.ntt.co.jp

This study was supported by
the Ministry of Internal Affairs and Communications of Japan.

Q: Will Calculation Complexity Be Explosion?

- The way of making single dimensional tree and compressing way is similar to Estan's way in [Automatically].
- So, number of nodes on compressed tree is limited,
- We can Search all non-overlap node combinations in a brute force way within realistic time and resource.



[Automatically]: C. Estan, S. Savage and G. Varghese, "Automatically Inferring Patterns of Resource Consumption in Network Traffic," *SIGCOMM, August 2003.*