Advancing Metrics on the Standards Track:

RFC 2679
Test Plan and Results

draft-morton-ippm-testplan-rfc2679-01

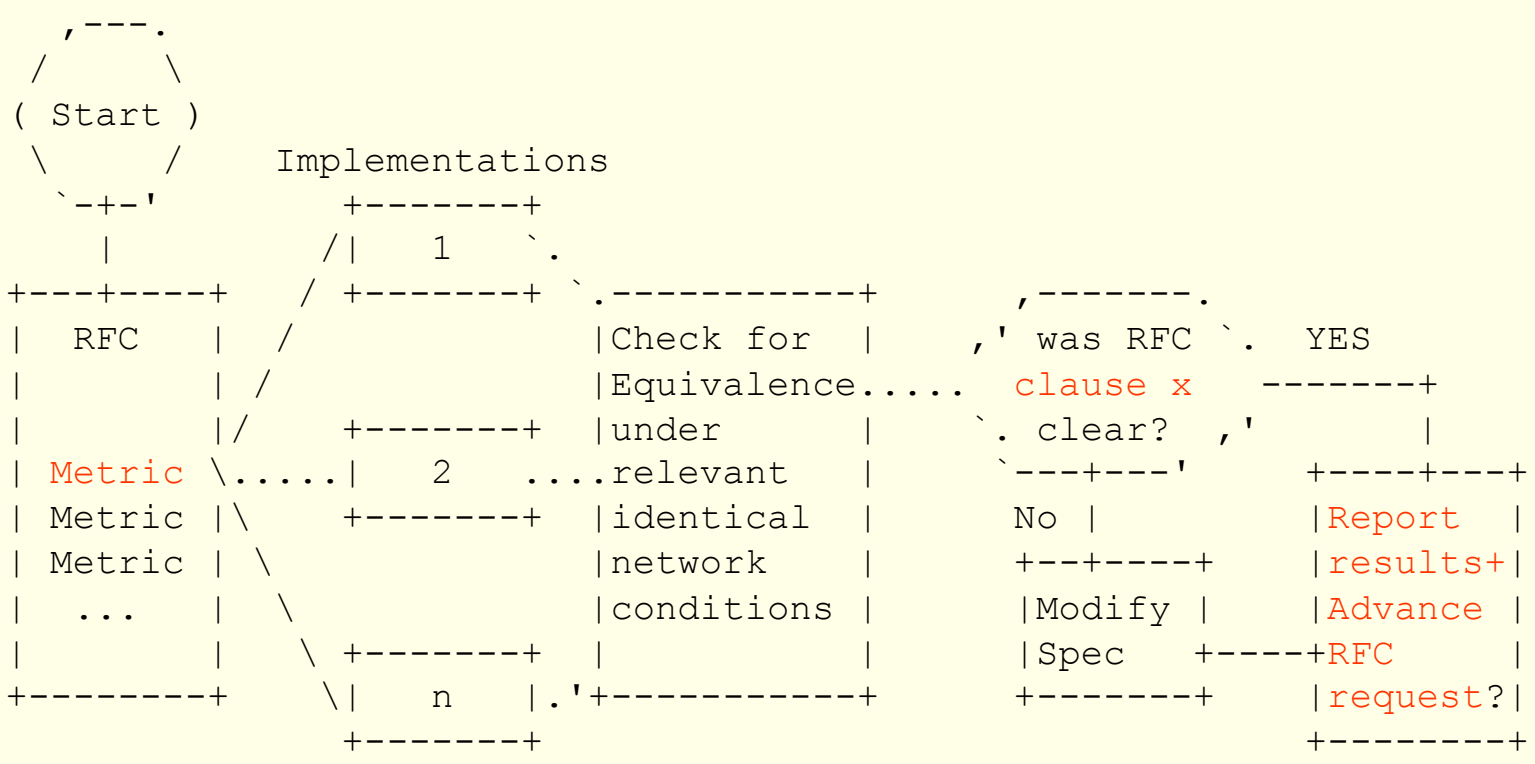
Len Ciavattone, Rüdiger Geib,
Al Morton, Matthias Wieser

July 2011

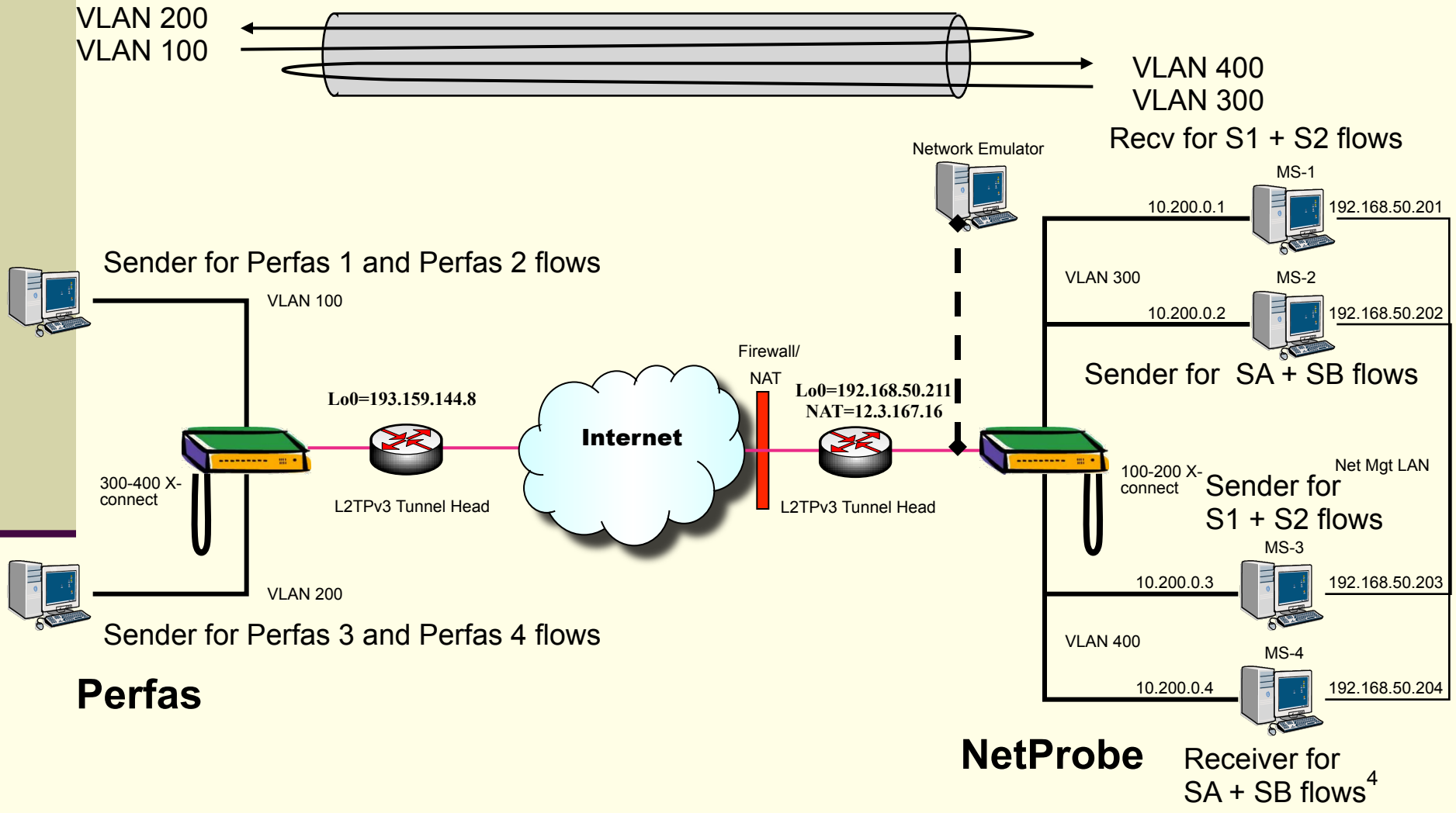
Outline

- Implement the Definition-centric metric advancement described in “metrictest” draft
- Test Plan Overview
  - Test Set-up and Specific Tests
- Test Results
- Summary and implications on the text of the revised RFC2679

Definition-Centric Process



Test Configuration



Tests in the Plan

- 6. Tests to evaluate RFC 2679 Specifications
  - 6.1. One-way Delay, ADK Sample Comparison – Same & Cross Implementations
  - 6.2. One-way Delay, Loss threshold,
  - 6.3. One-way Delay, First-bit to Last bit,
  - 6.4. One-way Delay, Difference Sample Metric
  - 6.5. Implementation of Statistics for One-way Delay

Section 6.1 One-way Delay, ADK Sample Comparisons (Same/Cross)

1. Configure tests on an L2TPv3 tunnel over a live network path.
2. Measure a sample of one-way delay singletons with 2 or more implementations, using identical options.
3. Measure a sample of one-way delay singletons with *four* instances of the *same* implementations,
   - connectivity differences SHOULD be the same as for the *cross* implementation tests.
4. Apply ADK comparison: same (see App C of metrictest)
5. Take coarsest confidence/resolution, or Section 5 Limits
6. Apply constant correction factors (Section 5)
7. Compare Cross-Implementation ADK for equivalence (samples come from same distribution)

Criteria for the Equivalence Threshold and Correction Factors

- Purpose: Evaluate Specification Clarity (using results implementations)
- For ADK comparison: cross-implementations
  - 0.95 confidence factor at 1ms resolution, or
  - The smallest confidence factor & res. of *same* Imp.
- A constant time accuracy error < +/-0.5ms MAY be removed from one Implementation before ADK or comparison of means
- A constant propagation delay error < +2ms MAY be removed from one Implementation …
  - (due to use of different sub-nets between the switch and measurement devices at each location)

Overview of Testing

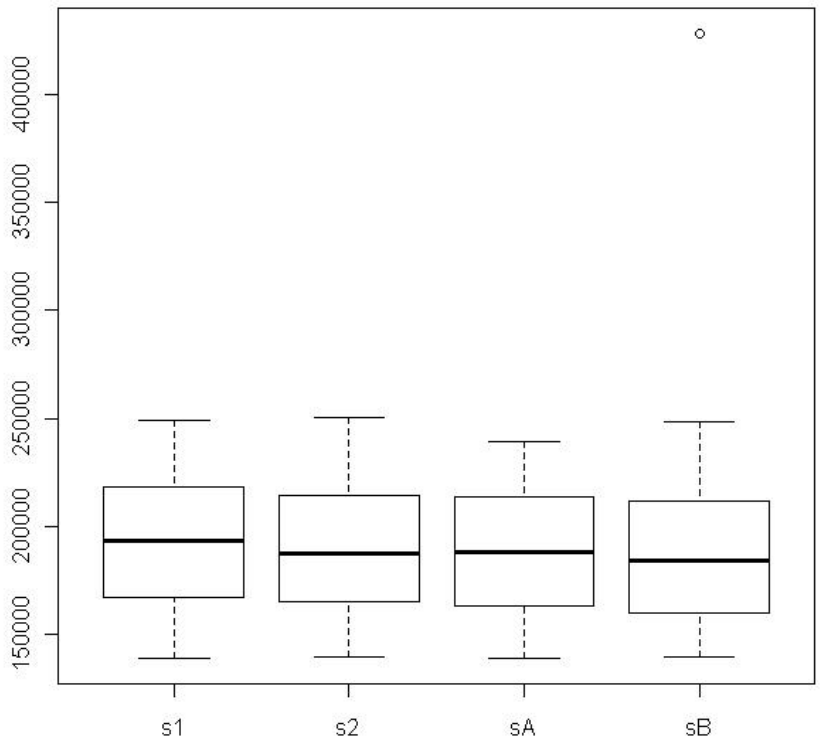
- 32 different experiments conducted from March 9 through May 2, 2011.
- Varied Packet size, Active sampling distribution, test duration, and other parameters (Type-P)
- Added Network Emulator “netem” and varied fixed and variable delay distirbutions
  - Also inserted loss in a limited number of experiments.

Overview of Testing (sample)

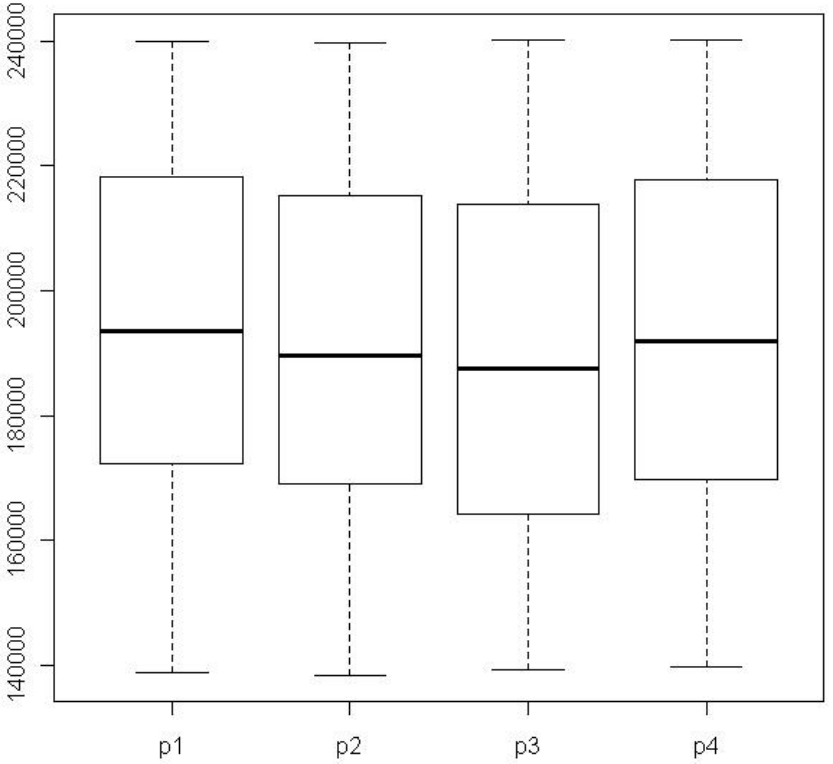
Date	Samp	Interval	Duration	Notes	ADK same	ADK cross
Mar 23	Poisson	1s	300s	Netem 10% Loss		
Mar 24	Periodic	1s	300s	Netem 100ms +/- 50ms delay		
Mar 24	Periodic	1s	300s	Netem 10% Loss		
Mar 28	Periodic	1s	300s	Netem 100ms		
Mar 29	Periodic (rand st.)	1s	300s	Netem 100ms +/- 50ms delay, 64 Byte	NP s12AB Per p1234	Pass combined
Apr 6	Periodic (rand st.)	1s	300s	Netem 100ms +/- 50ms delay, 340 Byte		
Apr 7	Periodic (rand st.)	1s	1200s	Netem 10% Loss		
Apr 12	Periodic (rand st.)	1s	300s	Netem 100ms, 500 Byte and 64 Byte comparison		

Summary of March 29 Tests
No correction factors used, 1usec res.

- NetProbe



- Perfas+



ADK tests – Glossary & Background

The ADK R-package returns some values and these require interpretation:

ti.obs is calculated, an observed value based on an ADK metric. The absolute ti.obs value must be less than or equal to the Critical Point.

The P-value or (P) in the following tables is a statistical test to bolster confidence in the result. It should be greater than or equal to $\alpha = 0,05$.

Critical Points for a confidence interval of 95% (or $\alpha = 0.05$)
For k = 2 samples, the Critical Point is 1.960
For k = 4 samples, the Critical Point is 1.915
For k = 9 samples, the Critical Point is 1.839
(Note, the ADK publication doesn't list a Critical Point for 8 samples, but it can be interpolated)

Green = ADK test passed, Red = ADK test failed

ADK for Mar 29 tests - NetProbe

ADK ti.obs (P)	NetProbe 1	NetProbe 2	NetProbe A
NetProbe 2	0.25 (0.28)		
NetProbe A	0.60 (0.19)	-0.80 (0.57)	
NetProbe B	2.64 (0.03)	0.07 (0.31)	-0.52 (0.48)

NetProbe ADK Results for same-implementation

Green = passed, Red = failed

ADK for Mar 29 tests – Perfas+

ADK ti.obs (P)	perfas 1	perfas 2	perfas 3
perfas 2	0.06 (0.32)		
perfas 3	1.09 (0.12)	0.37 (0.24)	
perfas 3	-0.81 (0.57)	-0.13 (0.37)	1.36 (0.09)

Perfas ADK Results for same-implementation

Green = passed, Red = failed

ADK for Mar 29 – Cross-Implementations

Null Hypothesis:
All samples within a data set come from a common distribution. The common distribution may change between data sets.

All NetProbe combined	ti.obs	P-value
not adj. for ties	0.64999	0.21355
adj. for ties	0.64833	0.21392
All Perfas combined		
not adj. for ties	0.55968	0.23442
adj. for ties	0.55840	0.23473
All Netprobe and Perfas combined		
not adj. for ties	0.85537	0.17967
adj. for ties	0.85329	0.18010

Green = passed, Red = failed

Other Results (details in the memo)

- Calibration – completed for both implementations
- Loss Threshold – available in post-processing for both implementations
  - Suggest revised text to allow this in RFC
- First bit – Last bit – issues with test design
  - Low speed links not available
  - Emulator interfaces found in Half-Duplex
  - Replace with descriptions of implementations
- Differential Delay – sufficiently accurate
- Delay Stats – drop Type-P-One-way-Delay-Inverse-Percentile in this RFC

Summary

- Two Implementations: NetProbe and Perfas+
- Test Plan for Key clauses of RFC 2679
  - the basis of Advance RFC Request
  - Criteria for Equivalence Threshold & correction factors
  - Adopt as a WG document?
- Experiments complete, key clauses of RFC2679 evaluated
  - Two revisions to the RFC suggested from this study

References

- R Development Core Team (2011), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
- Scholz F.W. and Stephens M.A. (1987), K-sample Anderson-Darling Tests, *Journal of the American Statistical Association,* **Vol 82, No. 399**, 918–924.

BACKUP

Backup Backup Backup

Percentiles of the ADK Criteria for various sample combinations (k= number of samples) [Table 1 of Scholz and Stevens]

m (k-1)	0.75 α=0.25	0.90 α=0.1	0.95 α=0.05	0.975 α=0.025	0.99 α=0.01
1	.326	1.225	1.960	2.719	3.752
2	.449	1.309	1.945	2.576	3.414
3	.498	1.324	1.915	2.493	3.246
4	.525	1.329	1.894	2.438	3.139

Criteria met when |t.obs| < ADK Criteria(%-tile of interest)

Also: P-value should be > α (rule of thumb)

Test Set-up Experiences

- Test bed set up may have to be described in more detail.
- We've worked with a single vendor.
- Selecting the proper Operation System took us one week (make sure support of L2TPv3 is a main purpose of that software).
- Connect the IPPM implementation to a switch and install a cable or internal U-turn on that switch. Maintain separate IEEE 802.1q logical VLAN connections when connecting the switch to the CPE which terminates the L2TPv3 tunnel.
- The CPE requires at least a route-able IP address as LB0 interface, if the L2TPv3 tunnel spans the Internet.
- The Ethernet Interface MUST be cross connected to the L2TPv3 tunnel in port mode.
- Terminate the L2TPv3 tunnel on the LB0 interface.
- Don't forget to configure firewalls and other middle boxes properly.

NetProbe 5.8.5

- Runs on Solaris (and Linux, occasionally)
- Pre-dates *WAMP, functionally similar
- Software-based packet generator
- Provides performance measurements including Loss, Delay, PDV, Reordering, Duplication, burst loss, etc. in post-processing on stored packet records

Section 6.2 – Loss Threshold

- See Section 3.5 of [RFC2679], 3rd bullet point and also Section 3.8.2 of [RFC2679].
- 1. configure a path with 1 sec one-way constant delay
- 2. measure (average) one-way delay with 2 or more implementations, using identical waiting time thresholds for loss set at 2 seconds
- 3. configure the path with 3 sec one-way delay (or change the delay while test is in progress, measurements in step 2)
- 4. repeat measurements
- 5. observe that the increase measured in step 4 caused all packets to be declared lost, and that all packets that arrive successfully in step 2 are assigned a valid one-way delay.

Section 6.3: First-bit to Last-bit

See Section 3.7.2 of [RFC2679], and Section 10.2 of [RFC2330].

- 1. configure a path with 1000 ms one-way constant delay, and ideally including a low-speed link (10-baseT, FD)
- 2. measure (average) one-way delay with 2 or more implementations, using identical options and equal size small packets (e.g., 44 octet IP payload)
- 3. maintain the same path with 1000 ms one-way delay
- 4. measure (average) one-way delay with 2 or more implementations, using identical options and equal size large packets (e.g., 480 octet IP payload)
- 5. observe that the increase measured in steps 2 and 4 is equivalent to the increase in ms expected due to the larger serialization time for each implementation. Most of the measurement errors in each system should cancel, if they are stationary.

Other Examples

- 6.4 One-way Delay, RFC 2679
  - This test is intended to evaluate measurements in sections 3 and 4 of [RFC2679].

    Average delays before/after 2 second increase

- 4. Error Calibration, RFC 2679
  - This is a simple check to determine if an implementation reports the error calibration as required in Section 4.8 of [RFC2679].