# Comparing Metrics Using -bradner-metricstest-00- A Small Study

Matt Zekauskas, matt@advanced.org

51st IETF Meeting

London

7 August 2000

# Goal

- Attempt to apply the method in draft-bradner-metricstest-00.txt, and the values suggested by Vern Paxson at the last IPPM meeting to get some practical experience

# Summary of Method

- Run two methods for a given metric on a randomized schedule
- May run simultaneously if appropriate
- Perhaps: method B's value falls within $2\sigma$ of A's value at least 90% [*] of the time. (If truly identical, expect 95%.)

# Factors Ignored In This Talk

- Randomized schedules
- Coverage for Type-P
- Comparing singleton values
- … part of why it's a "small" study
- … another part is that it is only for a few days, and also that the analysis isn't complete
- Ignoring BMWG-like questions, focusing on IPPM questions

# Data: Surveyor

- 12 byte UDP packets
- Poisson Schedule
- Averages usually O(1 second)
- Gives up if GPS lock lost, even momentarily (implies holes in the data)

# Data: Test-Traffic

- A RIPE-NCC project
- 100 byte packets
- Poisson schedule
- Average usually O(20 seconds)

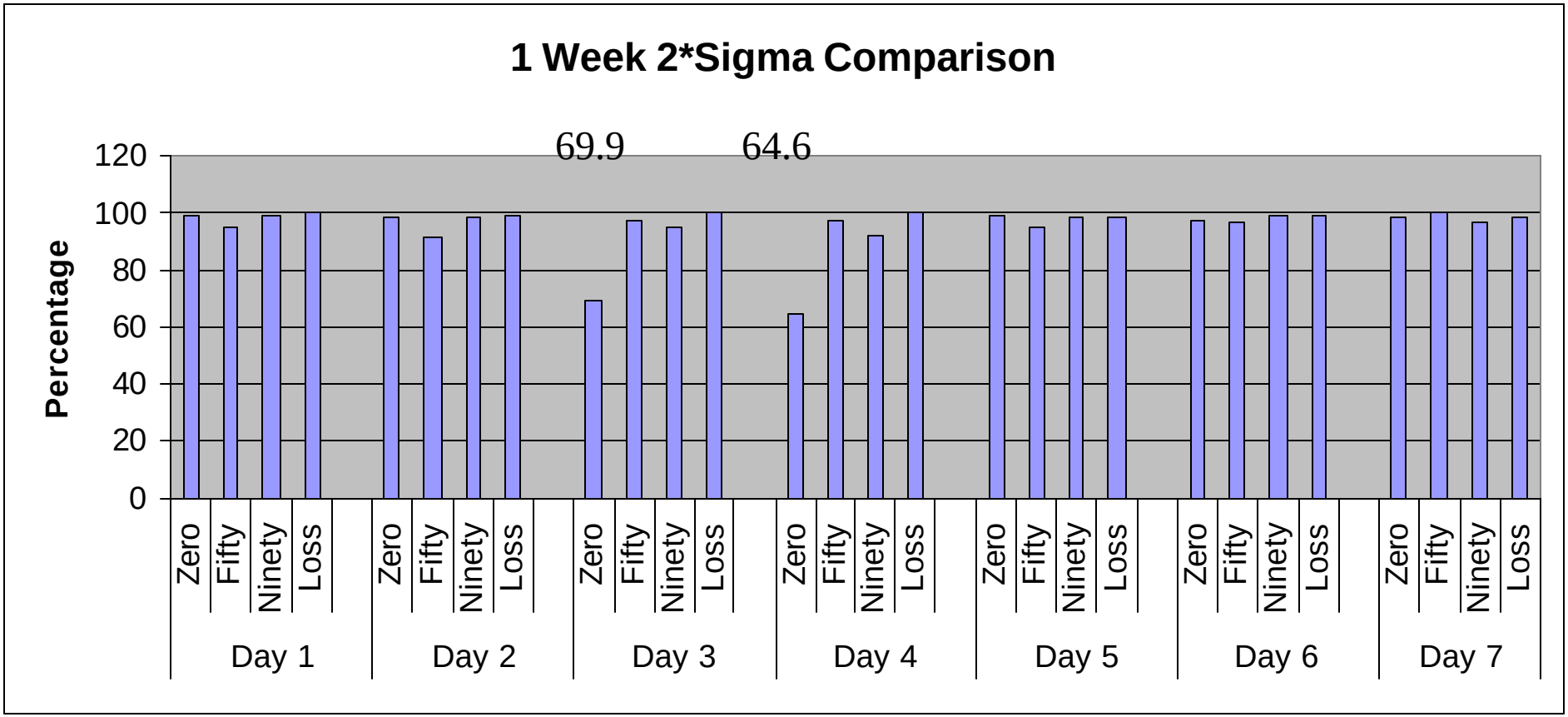
# The Experiment

- Change Surveyor to use 100 byte packets, 20 second average
- [Data is NOT the same]
- Focus on the Amsterdam – Armonk, NY path, covered by both RIPE-NCC Test Traffic and Surveyor. Machines are on the same Ethernet segments
- Both PC-class machines. But different PC's, GPS hardware, and software.

# Performing the Comparison

- What about those infinite values?
  - Ignored for the purpose of computing μ, σ
  - When comparing, the same iff both ∞
- What about missing values?
  - For any time range where there is missing data from one “method”, delete values from other method, not used at all either for μ, ó or comparison
- What are we using to compute μ and σ?
  - Min, 50$^{th}$ percentile, 90$^{th}$ percentile of 5 minute chunks, μ, σ computed separately for each

# 14-June to 20-June-2001



# Anomaly Analysis

- On Day 3 and Day 4 (the weekend), the standard deviation is less than the expected error in the Surveyor’s results (33 microseconds and 30 microseconds compared with 100 microseconds at current measurement load)
- Others: 495, 1217, 285, 188, 6576 µs
- Data range: 40 to 46 typ., max ~820 ms

# Comparing Singletons

- How decide which to compare, given that methods are using independent Poisson processes to schedule the measurements?

# Observations

- Clearly lots of room to tailor to your particular situation
- Must understand the goals of the comparison
- Singleton comparison hard, given time sensitivity, and randomness inherent in Poisson process