Towards a Methodology for Experimental Evaluation in Low-Power Wireless Networking

Romain Jacob
Usman Raza
Lothar Thiele

Carlo Alberto Boano
Marco Zimmerling

CPS-IoTBench Workshop
April 15, 2019

Includes material from Hanspeter Schmid and Alex Huber
“We need a benchmark for IoT networking.”
“We need a benchmark for IoT networking.”

⇔ Comparing performance
"We need a benchmark for IoT networking."

⇔ Comparing performance

⇒ Repeatable experiments
“We need a benchmark for IoT networking.”

⇔ Comparing performance

⇒ Repeatable experiments

⇒ Formalize the experimental methodology
The RF environment affects performance of low-power protocol in unpredictable ways.*

Real RF environment cannot be controlled.

Performance variability is expected.

* Either unfeasible or unpractical to model
How do you handle variability?

"Repeat your experiment." Easy, right?
How do you handle variability?

“Repeat your experiment.” Easy, right?

Killer questions

- How long should be your experiment?
- How many times should you repeat it?
How do you handle variability?

“Repeat your experiment.” Easy, right?

How long should be your experiment?
How many times should you repeat it?

“Run many long tests.” Not so easy...
Let us assume you ran “many long tests” on 1 M samples of XYZ spread over a large period of time.

How do you synthesize your results?

“Use statistics.”
Let us assume you ran “many long tests”

1 M samples of XYZ spread over a large period of time

How do you synthesize your results?

“Use statistics.”

Literally

A piece of data obtained from a large quantity of data

Mean, median, standard deviation, etc.
Beware!

Descriptive statistics ≠ Predictive statistics

What is the collected data like

What the collected data allows to infer about future/other data (unknown)
Beware!

Descriptive statistics ≠ Predictive statistics

What is the collected data like

What the collected data allows to infer about future/other data (unknown)

The “interesting case”
Descriptive statistics compare the samples

Sample A

Sample B

Conclusion       Sample A is “better” than sample B.
Predictive statistics (aims to) compare the underlying distributions

Sample A

Dist. A

Conclusion

A is “better” than B.

If one sample A and B, then likely sample A is “better” than sample B.
Descriptive statistics

Sample A is “better” than sample B.

Predictive statistics

If one sample A and B, then likely sample A is “better” than sample B.

Much stronger statement but also harder to make
Tendency | Mean |
----|----|
Variability | Standard deviation |

Prediction?
If we use mean and standard deviation in this context, we make two mistakes

First error: The standard deviation of the sample is not the standard deviation of the underlying distribution.

Use confidence intervals
Confidence interval (CI)?

Informally

A numerical interval in which lies the true value (which you don’t known) of some parameter with some probability (or confidence level)

Example

\([a, b]\) is a 95% CI for the mean of \(x\)

\(\iff\) The probability than the true mean value of \(x\) is included in \([a, b]\) is larger or equal to 95%
If we use mean and standard deviation in this context, we make two mistakes

First error  The standard deviation of the sample is not the standard deviation of the underlying distribution.

Use confidence intervals

Second error  The underlying distribution is not normal!

The standard deviation does not help making predictions
Use non-parametric statistics
Towards a Methodology for Experimental Evaluation in Low-Power Wireless Networking

Know your data
Use non-parametric statistics

Formalizing low-power wireless experimental evaluation
Performance measurements in computer science are typically **not normally distributed**
Performance measurements in computer science are typically not normally distributed

Use non-parametric statistics based on distribution percentiles

$\begin{align*}
& \text{p-th percentile or } Pp \\
& p\% \quad \text{of the distribution is below} \\
& (1 - p)\% \quad \text{of the distribution is above}
\end{align*}$

Great for predictive statistics
Percentiles are powerful predictive statistics

- **Simple to use**: Can compute CI for any percentile with any confidence
- **Distribution independent**: Estimates are valid regardless of the underlying distribution
- **Robust**: Estimates are not skewed by outliers
Confidence intervals

\[ P\{x_m \leq M \leq x_{N-m+1}\} = 1 - 2 \sum_{k=0}^{m-1} \binom{N}{k} \frac{1}{2^N} \]

\[ P\{x_m \leq P_p\} = \frac{1}{2^N} \sum_{k=0}^{N-1} \binom{N}{k} p^k(1-p)^{N-k} \]
Confidence intervals

\[ P \left\{ x_m \leq M \leq x_{N-m+1} \right\} = 1 - 2 \sum_{k=0}^{m-1} \binom{N}{k} \frac{1}{2^N} \]

\[ P \left\{ x_m \leq P_p \right\} = \]

\[ P \left\{ x_{N-m+1} \geq P_{1-p} \right\} = 1 - \sum_{k=0}^{m-1} \binom{N}{k} p^k (1 - p)^{N-k} \]
Hypothesis: Samples are i.i.d.
Hypothesis: Samples are i.i.d.
Hypothesis: Samples are i.i.d.
Binomial distribution

\[ P \{ x_k \leq p \leq x_{k+1} \} = \binom{N}{k} p^k (1 - p)^{N-k} \]
Confidence intervals

\[ P \{x_m \leq M \leq x_{N-m+1}\} = 1 - 2 \sum_{k=0}^{m-1} \binom{N}{k} \frac{1}{2^N} \]

\[ P \{x_m \leq P_p\} = 1 - m \sum_{k=0}^{m-1} \binom{N}{k} p^k (1-p)^{N-k} \]

\[ P \{x_{N-m+1} \geq P_{1-p}\} = 1 - m \sum_{k=0}^{m-1} \binom{N}{k} p^k (1-p)^{N-k} \]
\[ P = 99.29\% \]

when 2 x 0 points are excluded
$P = 92.97\%$

when 2 x 1 points are excluded
$M \diamond \diamond \diamond \diamond \diamond \diamond \diamond$ 

$P = 71.09\%$

when 2 x 2 points are excluded
\[ M = 27,34\% \]

when 2 x 3 points are excluded
Percentiles are powerful predictive statistics

- Simple to use
- Distribution independent
- Robust

Can compute CI for any percentile with any confidence

Estimates are valid regardless of the underlying distribution

Estimates are not skewed by outliers
Percentiles are powerful predictive statistics

Simple to use

Can compute CI for any percentile with any confidence

Distribution independent

Estimates are valid regardless of the underlying distribution

Robust

Estimates are not skewed by outliers
For any confidence $c$, any percentile $p$,

$$N \geq \frac{\log(1 - c)}{\log(1 - p)}$$

Example

$c = 0.95$
$p = 0.01$ or 1-th percentile

$\Rightarrow N \geq 299$
We can derive the **minimal number of samples** required for estimating any percentile with any confidence.

Thus
We can derive the minimal number of samples required for estimating any percentile with any confidence.

So now How long should be your experiment? How many times should you repeat it? Can be answered rationally.
Towards a Methodology for Experimental Evaluation in Low-Power Wireless Networking

Know your data
Use non-parametric statistics

Formalizing low-power wireless experimental evaluation
Test Configuration → Experiments → Performance Results

Comm. Protocol A → Experiments → Performance Results

Better than B?

Yes →

No →
Test Configuration → Experiments → Performance Results → Better than B? → Yes/No
Test Configuration

Comm. Protocol A

Experiments

Performance Results

Better than B?

Repeateble?

Yes

Yes

No

No
Which metric to compute?

Experiments

1. Experiment 1
   Raw Data → Analysis → Processed Data

2. Experiment 2
   Raw Data → Analysis → Processed Data

... (N experiments)

Test Configuration
Comm. Protocol

Performance Results

Synthesis
Which raw data to collect?

1. Which metric to compute?

2. Which raw data to collect?
Experiments

Which metric to compute?

Raw Data

Analysis

Processed Data

Experiment 1

Experiment 2

::

Experiment N

Raw Data

Analysis

Processed Data

Synthesis

Performance Results

Which raw data to collected?

How long should the experiment be?

Comm. Protocol
Which raw data to collect?

How long should the experiment be?

Which metric to compute?

How to synthesize performance results?
2nd Workshop on Benchmarking Cyber-Physical Networks and Internet of Things (CPS-IoTBench)
April 15, 2019 | Montréal | Canada

Which raw data to collect?

How many experiments?

How long should the experiment be?

Which metric to compute?

How to synthesize performance results?

Experiments

Raw Data

Analysis

Processed Data

Performance Results

Synthesis

Test Configuration

Comm. Protocol

1. Which metric to compute?

2. Experiment 1

3. Experiment 2

4. How to synthesize performance results?

5. Experiment N
Case study – Periodic data collection

14 source nodes
200 payloads per source
2 Bytes per payload
10 payload per second
Periodic release, asynchronous

First payload released after 10s
Test stops 10s after last payload is released
Select the “metrics” based on the purpose of the evaluation

<table>
<thead>
<tr>
<th>Performance dimensions</th>
<th>Reliability</th>
<th>Average</th>
<th>Extremal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Latency</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Energy efficiency</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radio duty-cycle</td>
</tr>
<tr>
<td>Current draw</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Median/percentiles</td>
</tr>
<tr>
<td>Max/Min</td>
</tr>
</tbody>
</table>
Select the “metrics” based on the purpose of the evaluation.

How many application payloads can one expect to successfully receive in one execution of the scenario?

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Metric</th>
<th>Measure</th>
<th>We are trying to predict future performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average reliability</td>
<td>PRR</td>
<td>Mean</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Median</td>
<td></td>
</tr>
</tbody>
</table>
Collect raw data with the finest granularity possible

PRR
Log all received payloads at the sink

Current draw
14400 samples/s 1 every ~7 μs
10 pA precision
Define the length of the experiment based on the scenario and the protocol

- Generally difficult
  - Correct approach depends on the protocol under test

- Easy case
  - If scenario is terminating and short, then run it in full
Define the length of the experiment based on the scenario and the protocol

Terminating

200 payload/source

+ Short

10 + 200/10 + 10 = 40s

⇒ Run in full

⇒ The uncertainty lies only in the variability across experiments
Use performance indicators based on confidence intervals

- **Average reliability**
  - Overall PRR
    - \( \forall \) experiment \( j \), \( x_j \)

  - 95% CI on the median PRR for all exp.
    - \( [x_m, x_{N-m+1}] \)

- **Use conservative bound**
  - \( x_m \)

---

**Received payloads**

\[
200 \times 14
\]
Perform sufficiently many experiments to obtain tight CI.

Intuition: Estimating average performance is easier than extremal performance.

Need more experiments to estimate a 95-th percentile than a median.
Perform sufficiently many experiments to obtain tight CI

Performance indicators based on 95% CI for the median

\[ N \geq 6 \]

Aim for tighter CI

\[ N = 20 \]

\[ 95\% \text{ CI is } [x_6, x_{15}] \]
Following the methodology enables **unambiguous** performance reports.
Following the methodology enables **unambiguous** performance reports.

<table>
<thead>
<tr>
<th>Protocol</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Energy</td>
<td>0.82</td>
<td>0.83</td>
<td>0.89</td>
<td>0.86</td>
<td>0.90</td>
<td>0.43</td>
<td>0.25</td>
</tr>
<tr>
<td>Worst-case Energy</td>
<td>0.67</td>
<td>0.44</td>
<td>0.82</td>
<td>0.18</td>
<td>0.52</td>
<td>0.27</td>
<td>0.19</td>
</tr>
<tr>
<td>Reliability</td>
<td>0.40</td>
<td>0.41</td>
<td>0.89</td>
<td>0.06</td>
<td>0.48</td>
<td>0.27</td>
<td>0.25</td>
</tr>
</tbody>
</table>

We are on good way...
"We need a benchmark for IoT networking."

⇔ Comparing performance

⇒ Repeatable experiments

⇒ Formalize the experimental methodology
Towards a Methodology for Experimental Evaluation in Low-Power Wireless Networking

Know your data
Use non-parametric statistics

Formalizing low-power wireless experimental evaluation
<table>
<thead>
<tr>
<th>Problem 1</th>
<th>Predictive statistics require i.i.d. measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not a given. This must be checked, not assumed.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Problem 2</th>
<th>What if the scenario is not terminating?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>We still don’t know how long one experiment should be.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Problem 3</th>
<th>To be comparable, results must be repeatable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>We still don’t know how to formalize repeatability in our context.</td>
</tr>
</tbody>
</table>

These are work-in-progress...
Towards a Methodology for Experimental Evaluation in Low-Power Wireless Networking

Romain Jacob
Usman Raza
Lothar Thiele

Carlo Alberto Boano
Marco Zimmerling

Includes material from Hanspeter Schmid and Alex Huber