Appendix: An Empirical Evaluation of the Energy and Performance Overhead of Monitoring Tools on Docker-based Systems

July 10, 2023

Table 1 reports the detailed statistics about Figure 1 in Section 5.1 of the paper. The energy efficiency across all monitoring tools ranges between 38,552 and 88,516 Joules. The coefficient of variation is between 21.3% and 26.8% and the standard deviation shows quite some variability of the data (13,453 globally), which is most probably coming from the difference among the blocks. Netdata is the most energy-efficient tool (mean 54,543 Joules), while Zipkin is the least one (mean 60,668 Joules).

Table 1: *Descriptive statistics of the energy efficiency in Joules (SD=standard deviation, CV=coefficient of variation)*

Evidence	Min.	Max.	Median	Mean	SD	CV
Baseline	39,368	72,117	51,396	53,755	11,461	21.3
ELK Stack	38,552	80,953	52,371	56,760	13,372	23.6
Netdata	39,827	78,334	49,340	54,543	12,342	22.6
Prometheus	41,246	80,115	49,464	55,046	12,485	22.7
Zipkin	40,976	88,516	54,796	60,668	16,248	26.8
Global	38,552	88,516	51,160	56,155	13,453	24.0

Tables 2, 3, 4, 5, and 6 report the detailed statistics about Figure 3 in Section 5.2 in the paper.

Table 2: *Descriptive statistics of the CPU usage % (SD=standard deviation, CV=coefficient of variation)*

Evidence	Min.	Max.	Median	Mean	SD	CV
Baseline	25.7	77.8	49.0	51.7	16.1	31.2
ELK Stack	26.7	69.4	47.2	49.0	13.7	28.0
Netdata	29.5	77.3	49.3	51.6	15.3	29.6
Prometheus	27.7	76.7	51.2	52.1	14.6	27.9
Zipkin	31.3	78.8	57.8	57.6	15.9	27.5
Global	25.7	78.8	50.5	52.4	15.3	29.2

Table 3: *Descriptive statistics of the CPU load average (SD=standard deviation, CV=coefficient of variation)*

Evidence	Min.	Max.	Median	Mean	SD	CV
Baseline	4.72	36.6	13.5	16.7	9.92	59.3
ELK Stack	6.11	34.3	13.8	17.3	9.15	53.0
Netdata	5.91	36.6	13.1	16.6	9.43	56.8
Prometheus	5.71	36.4	14.1	16.6	8.90	53.7
Zipkin	6.83	34.7	15.2	18.4	9.7	52.7
Global	4.72	36.6	13.9	17.1	9.41	55.0

Table 4: *Descriptive statistics of the RAM usage % (SD=standard deviation, CV=coefficient of variation)*

Evidence	Min.	Max.	Median	Mean	SD	CV
Baseline	52.3	81.0	64.0	65.1	6.73	10.34
ELK Stack	99.2	99.3	99.2	99.2	0.01	0.01
Netdata	54.3	84.9	68.4	68.3	8.15	11.93
Prometheus	52.7	83.6	67.5	67.1	8.29	12.36
Zipkin	57.3	87.9	70.8	71.0	7.01	9.87
Global	52.3	99.3	70.7	74.1	14.4	19.4

Table 5: *Descriptive statistics of the execution time (s) (SD=standard deviation, CV=coefficient of variation)*

Evidence	Min.	Max.	Median	Mean	SD	CV
Baseline	623	980	807	813	81.5	10.0
ELK Stack	615	1,050	820	852	111.4	13.08
Netdata	700	1,027	780	812	82.4	10.14
Prometheus	602	1,065	781	814	93.2	11.45
Zipkin	617	1,105	805	855	130.6	15.3
Global	602	1.105	804	829	103	12.4

Table 7a reports the p-values and η^2 results per variable of the Kruskal-Wallis test. The table refers to Section 5.2 of the paper.

Tables 8, 9, 10, 11, and 12 report the results of the Wilcoxon test for each frequency (F) and workload (W) combination of treatments. Table 8 reports that for ELK stack and Zipkin there is a significant impact of monitoring tools on CPU usage. The p-values for 4 blocks allow to reject the null hypothesis that the median difference between the baseline and ELK stack is zero. Similarly, the p-values for 6 blocks allow to reject the null hypothesis that the median difference between the baseline and Zipkin is zero. Table 9 reports that for Zipkin and Netdata there is at least one block where the impact on CPU load is significant. Table 10 reports that, except for Netdata, there is statistical significance for every tool, for at least one block. Table 11 reports that there is no statistical significance for network traffic. Finally, Table 12 reports that in case of ELK stack and Zipkin there is statistical significance for the execution time and Cliff's delta estimates show large effect size.

The scatter plot in Figure 2 refers to the end of Section 5.2 of the paper, it plot the correlation of each variable with energy efficiency

Figure 1 details dependent variables across all combinations of frequency and workload treatments. The RAM usage plot clearly shows a very high RAM utilization percentage, close to 100%. Hence monitoring tools can

Table 6: *Descriptive statistics of the network traffic (SD=standard deviation, CV=coefficient of variation)*

Evidence	Min.	Max.	Median	Mean	SD	CV
Baseline	444,920	2,430,224	1,012,841	1,227,166	730,416	59.5
ELK Stack	442,106	2,392,325	972,469	1,239,860	749,250	60.4
Netdata	447,636	2,377,342	987,881	1,242,516	746,946	60.1
Prometheus	457,694	2,387,461	974,433	1,242,673	750,686	60.4
Zipkin	458,973	2,410,738	1,025,735	1,274,329	766,754	60.2
Global	442,106	2,430,224	1e+06	1,245,309	745,720	59.9

Table 7: Results of the Kruskal-Wallis test, for each frequency (F) and workload (W) combination of treatments (block). Boldface text denotes a significant difference ($\alpha = .05$)

(a) CPU usage

Block	p-value	η^2	η^2 interpretation
F Low, W Low	**0.000955**	0.324	large
F Low, W Medium	**0.0345**	0.142	large
F Low, W High	**4.73e-05**	0.47	large
F Medium, W Low	**0.0015**	0.301	large
F Medium, W Medium	**6.52e-05**	0.454	large
F Medium, W High	**0.000277**	0.384	large
F High, W Low	**0.00036**	0.372	large
F High, W Medium	**0.000697**	0.339	large
F High, W High	**1.64e-06**	0.63	large

(b) CPU load

Block	p-value	η^2	η^2 interpretation
F Low, W Low	0.0528	0.119	moderate
F Low, W Medium	0.474	-0.0106	small
F Low, W High	0.106	0.0807	moderate
F Medium, W Low	0.0538	0.118	moderate
F Medium, W Medium	**0.0153**	0.184	large
F Medium, W High	0.287	0.0222	small
F High, W Low	**0.00173**	0.294	large
F High, W Medium	**0.0382**	0.136	moderate
F High, W High	**0.00728**	0.222	large

(c) Ram usage

Block	p-value	η^2	η^2 interpretation
F Low, W Low	**5.7e-05**	0.461	large
F Low, W Medium	**3.09e-05**	0.49	large
F Low, W High	**3.59e-05**	0.483	large
F Medium, W Low	**2.34e-05**	0.503	large
F Medium, W Medium	**4.19e-05**	0.475	large
F Medium, W High	**8.77e-05**	0.44	large
F High, W Low	**1.22e-05**	0.535	large
F High, W Medium	**2.49e-05**	0.5	large
F High, W High	**9.85e-06**	0.545	large

(d) Network traffic

Block	p-value	η^2	η^2 interpretation
F Low, W Low	0.119	0.0744	moderate
F Low, W Medium	**0.018**	0.176	large
F Low, W High	0.804	-0.0528	small
F Medium, W Low	0.348	0.0101	small
F Medium, W Medium	0.504	-0.0148	small
F Medium, W High	0.105	0.081	moderate
F High, W Low	0.154	0.0594	small
F High, W Medium	0.865	-0.0605	moderate
F High, W High	**0.00126**	0.31	large

(e) Execution time

Block	p-value	η^2	η^2 interpretation
F Low, W Low	0.109	0.0792	moderate
F Low, W Medium	0.436	-0.00486	small
F Low, W High	**8.59e-06**	0.551	large
F Medium, W Low	**0.0317**	0.146	large
F Medium, W Medium	0.47	-0.01	small
F Medium, W High	**6.08e-06**	0.568	large
F High, W Low	0.0903	0.0896	moderate
F High, W Medium	0.201	0.0439	small
F High, W High	**3.24e-07**	0.706	large

influence RAM usage, under specific frequency and workload conditions.

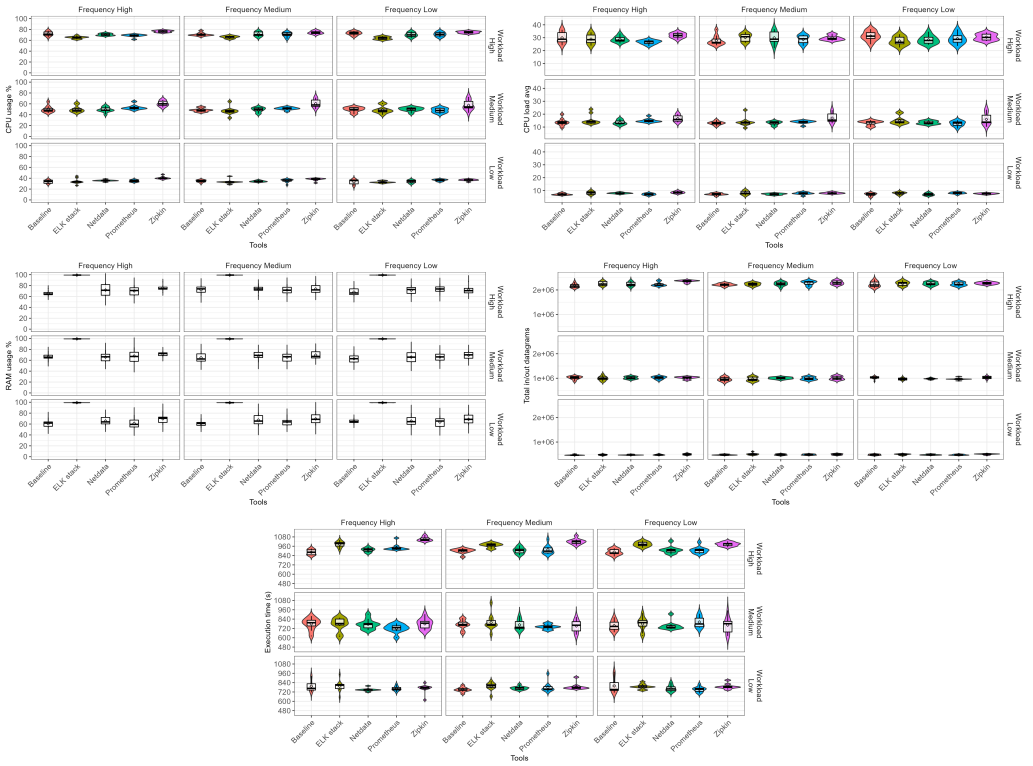

Figure 1: Dependent variables across all combinations of frequency and workload treatments

Table 8: *CPU usage results of the Wilcoxon test, for each frequency (F) and workload (W) combination of treatments (block) (statistically significant p-values are shown in bold)*

Tool	**Block**	**p-value**	**Cliff's δ**	**δ interpretation**
ELK stack	F Low, W Low	0.301	0.32	small
	F Low W Medium	0.591	0.20	small
	F Low, W High	**0.002**	0.96	large
	F Medium, W Low	**0.078**	0.58	large
	F Medium, W Medium	0.353	0.34	medium
	F Medium, W High	**0.004**	0.90	large
	F High, W Low	0.791	0.08	negligible
	F High, W Medium	1.000	0.00	negligible
	F High, W High	**0.009**	-0.66	large
Netdata	F Low, W Low	0.910	0.04	negligible
	F Low, W Medium	0.970	0.02	negligible
	F Low, W High	0.341	0.40	medium
	F Medium, W Low	0.427	0.22	small
	F Medium, W Medium	0.623	-0.14	negligible
	F Medium, W High	0.970	-0.12	negligible
	F High, W Low	0.651	-0.24	small
	F High, W Medium	1.000	-0.06	negligible
	F High, W High	0.473	0.20	small
Prometheus	F Low, W Low	0.301	-0.42	medium
	F Low, W Medium	0.575	0.26	small
	F Low, W High	0.341	0.30	small
	F Medium, W Low	0.202	-0.38	medium
	F Medium, W Medium	0.094	-0.56	large
	F Medium, W High	0.970	0.02	negligible
	F High, W Low	0.651	-0.20	small
	F High, W Medium	0.160	-0.50	large
	F High, W High	0.09	0.52	large
Zipkin	F Low, W Low	0.301	-0.50	large
	F Low, W Medium	0.078	-0.58	large
	F Low, W High	0.341	-0.30	small
	F Medium, W Low	**0.036**	-0.72	large
	F Medium, W Medium	**0.002**	-0.96	large
	F Medium, W High	**0.023**	-0.70	large
	F High, W Low	**0.001**	-0.98	large
	F High, W Medium	**0.011**	-0.82	large
	F High, W High	**0.035**	-0.66	large

Table 9: *CPU load results of the Wilcoxon test, for each frequency (F) and workload (W) combination of treatments (block) (statistically significant p-values are shown in bold)*

Tool	**Block**	**p-value**	**Cliff's δ**	**δ interpretation**
ELK stack	F Low, W Low	0.135	-5.20e-01	large
	F Low W Medium	0.779	-0.20	small
	F Low, W High	0.078	0.58	large
	F Medium, W Low	0.265	-0.34	medium
	F Medium, W Medium	0.791	-0.12	negligible
	F Medium, W High	0.107	-0.50	large
	F High, W Low	0.090	-0.52	large
	F High, W Medium	0.481	-0.24	small
	F High, W High	0.714	0.16	small
Netdata	F Low, W Low	1.000	-4.44e-17	negligible
	F Low, W Medium	0.910	-0.04	negligible
	F Low, W High	0.173	0.44	medium
	F Medium, W Low	0.970	-0.02	negligible
	F Medium, W Medium	0.791	-0.08	negligible
	F Medium, W High	0.301	-0.32	small
	F High, W Low	**0.043**	-0.64	large
	F High, W Medium	0.970	0.02	negligible
	F High, W High	0.970	0.02	negligible
Prometheus	F Low, W Low	0.135	-5.40e-01	large
	F Low, W Medium	0.779	0.18	small
	F Low, W High	0.341	0.30	small
	F Medium, W Low	0.233	-0.44	medium
	F Medium, W Medium	0.126	-0.48	large
	F Medium, W High	0.473	-0.20	small
	F High, W Low	0.970	-0.02	negligible
	F High, W Medium	0.260	-0.44	medium
	F High, W High	0.222	0.46	medium
Zipkin	F Low, W Low	0.384	-2.80e-01	small
	F Low, W Medium	0.779	-0.14	negligible
	F Low, W High	0.623	0.14	negligible
	F Medium, W Low	0.106	-0.62	large
	F Medium, W Medium	**0.014**	-0.74	large
	F Medium, W High	0.107	-0.54	large
	F High, W Low	**0.011**	-0.82	large
	F High, W Medium	0.128	-0.60	large
	F High, W High	0.353	-0.34	medium

Table 10: *RAM usage results of the Wilcoxon test, for each frequency (F) and workload (W) combination of treatments (block) (statistically significant p-values are shown in bold)*

Tool	**Block**	**p-value**	**Cliff's δ**	**δ interpretation**
ELK stack	F Low, W Low	**9.15e-04**	-1.0	large
	F Low W Medium	**9.15e-04**	-1.0	large
	F Low, W High	**9.15e-04**	-1.0	large
	F Medium, W Low	**9.15e-04**	-1.0	large
	F Medium, W Medium	**9.15e-04**	-1.0	large
	F Medium, W High	**9.15e-04**	-1.0	large
	F High, W Low	**9.15e-04**	-1.0	large
	F High, W Medium	**9.15e-04**	-1.0	large
	F High, W High	**9.15e-04**	-1.0	large
Netdata	F Low, W Low	0.910	0.04	negligible
	F Low, W Medium	0.591	-0.20	small
	F Low, W High	0.341	-0.32	small
	F Medium, W Low	0.232	-0.36	medium
	F Medium, W Medium	0.642	-0.24	small
	F Medium, W High	0.970	-0.06	negligible
	F High, W Low	0.232	-0.40	medium
	F High, W Medium	0.970	-0.02	negligible
	F High, W High	0.307	-0.28	small
Prometheus	F Low, W Low	0.910	0.06	negligible
	F Low, W Medium	0.591	-0.24	small
	F Low, W High	0.189	-0.48	large
	F Medium, W Low	0.307	-0.28	small
	F Medium, W Medium	0.970	-0.02	negligible
	F Medium, W High	0.970	0.02	negligible
	F High, W Low	0.970	0.02	negligible
	F High, W Medium	0.970	-0.06	negligible
	F High, W High	**6e-04**	-0.34	medium
Zipkin	F Low, W Low	0.642	-0.24	small
	F Low, W Medium	0.113	-0.54	large
	F Low, W High	0.341	-0.30	small
	F Medium, W Low	0.09	-0.54	large
	F Medium, W Medium	0.222	-0.46	medium
	F Medium, W High	0.970	-0.10	negligible
	F High, W Low	0.094	-0.56	large
	F High, W Medium	0.053	-0.62	large
	F High, W High	**6e-04**	-0.82	large

Table 11: *Network traffic results of the Wilcoxon test, for each frequency (F) and workload (W) combination of treatments (block) (statistically significant p-values are shown in bold)*

Tool	**Block**	**p-value**	**Cliff's δ**	**δ interpretation**
ELK stack	F Low, W Low	0.862	-0.26	small
	F Low W Medium	0.067	0.52	large
	F Low, W High	0.623	-0.14	negligible
	F Medium, W Low	0.445	-0.46	medium
	F Medium, W Medium	0.970	-0.02	negligible
	F Medium, W High	0.521	-0.18	small
	F High, W Low	0.714	-0.16	small
	F High, W Medium	0.970	0.22	small
	F High, W High	0.113	-0.54	large
Netdata	F Low, W Low	0.910	-0.04	negligible
	F Low, W Medium	**0.043**	0.64	large
	F Low, W High	0.534	-0.24	small
	F Medium, W Low	0.521	-0.20	small
	F Medium, W Medium	0.530	-0.38	medium
	F Medium, W High	0.384	-0.28	small
	F High, W Low	0.714	-0.18	small
	F High, W Medium	0.970	0.08	negligible
	F High, W High	0.232	-0.36	medium
Prometheus	F Low, W Low	0.9107	0.06	negligible
	F Low, W Medium	**0.043**	0.62	large
	F Low, W High	0.534	-0.22	small
	F Medium, W Low	0.521	-0.24	small
	F Medium, W Medium	0.642	-0.24	small
	F Medium, W High	0.160	-0.50	large
	F High, W Low	0.465	-0.36	medium
	F High, W Medium	0.970	0.02	negligible
	F High, W High	0.173	-0.44	medium
Zipkin	F Low, W Low	0.378	-0.48	large
	F Low, W Medium	0.623	-0.14	negligible
	F Low, W High	0.534	-0.28	small
	F Medium, W Low	0.521	-0.34	medium
	F Medium, W Medium	0.530	-0.34	medium
	F Medium, W High	0.128	-0.60	large
	F High, W Low	0.086	-0.64	large
	F High, W Medium	0.970	0.14	negligible
	F High, W High	**0.002**	-0.94	large

Table 12: *Execution time results of the Wilcoxon test for each frequency (F) and workload (W) combination of treatments (block) (statistically significant p-values are shown in bold)*

Tool	**Block**	**p-value**	**Cliff's δ**	**δ interpretation**
ELK stack	F Low, W Low	0.640	-0.24	small
	F Low W Medium	0.97	-0.18	small
	F Low, W High	**4.58e-03**	-1.00	large
	F Medium, W Low	0.064	-0.66	large
	F Medium, W Medium	0.970	-0.02	negligible
	F Medium, W High	**1e-03**	-0.94	large
	F High, W Low	0.705	-0.15	small
	F High, W Medium	1	-1.00e-02	negligible
	F High, W High	**1e-03**	-0.94	large
Netdata	F Low, W Low	0.910	0.04	negligible
	F Low, W Medium	0.97	-0.07	negligible
	F Low, W High	0.361	-0.29	small
	F Medium, W Low	0.125	-0.48	large
	F Medium, W Medium	0.917	0.10	negligible
	F Medium, W High	0.970	-0.07	negligible
	F High, W Low	0.705	0.29	small
	F High, W Medium	0.868	1.80e-01	small
	F High, W High	0.064	-0.50	large
Prometheus	F Low, W Low	0.910	0.06	negligible
	F Low, W Medium	0.70	-0.40	medium
	F Low, W High	0.364	-0.25	small
	F Medium, W Low	0.283	-0.33	medium
	F Medium, W Medium	0.530	0.40	medium
	F Medium, W High	0.970	-0.02	negligible
	F High, W Low	0.705	0.13	negligible
	F High, W Medium	0.246	5.30e-01	large
	F High, W High	**3e-03**	-0.83	large
Zipkin	F Low, W Low	0.495	-0.36	medium
	F Low, W Medium	0.97	-0.02	negligible
	F Low, W High	**4.58e-04**	-1.00	large
	F Medium, W Low	0.064	-0.60	large
	F Medium, W Medium	0.917	0.14	negligible
	F Medium, W High	**9.05e-04**	-1.00	large
	F High, W Low	0.705	-0.11	negligible
	F High, W Medium	1.000	4.44e-17	negligible
	F High, W High	**9.1e-04**	-1.00	large

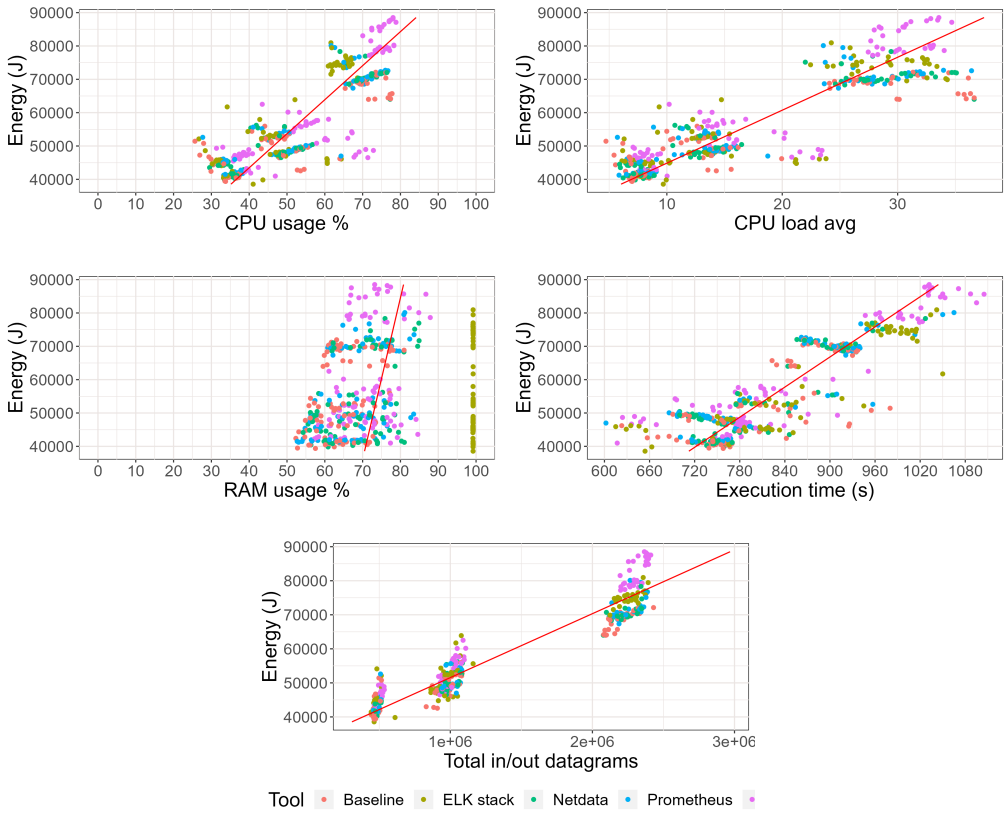

Figure 2: Correlation with energy efficiency