Final Q&A session

Ivano Malavolta

VU VRIJE UNIVERSITEIT AMSTERDAM

LOOKING FURTHER

Q1

Are run tables usually separate when using blocking factors? If so, why aren't these just split up into multiple research papers?

I initially had dataset size (one dataset that's larger and the other smaller) as a co-factor, but now I am thinking if it should be a blocking factor. If I change it to blocking factor, should I have two run tables (one for smaller dataset and the other for larger?

Blocking factors in practice?

Answers

The run table is typically one since you need to randomize the order of the runs also based on the blocking factor to avoid bias.

The decision between a co-factor and a blocking factor is primarily relevant for you during the data analysis phase, not during the execution of the experiment.

Blocking factor in practice: see next slide

Example (ICSOC 2024)

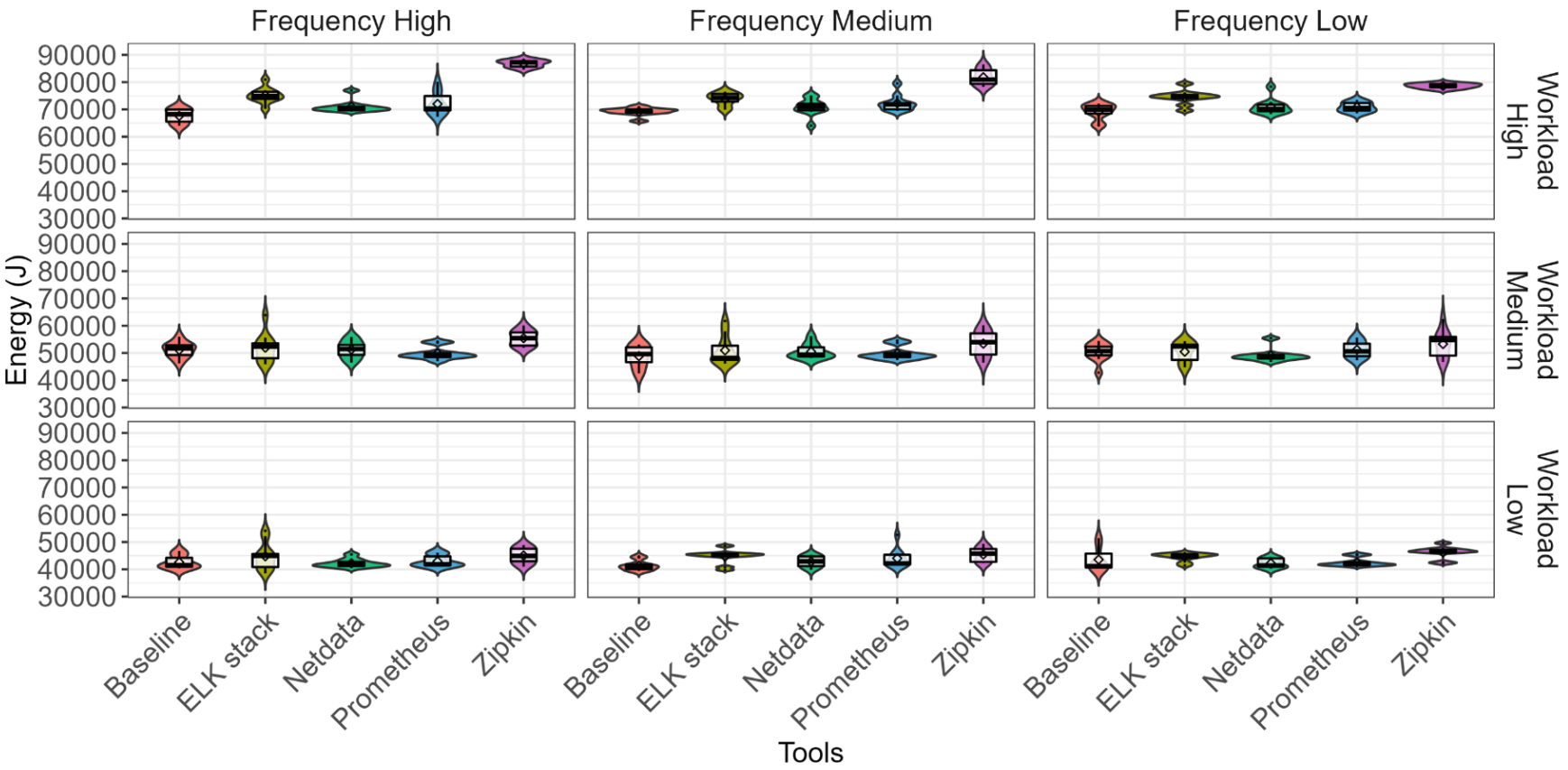
RQ3 Experiment Variables and Subjects Selection

RQ3: What is the runtime overhead of monitoring tools for Docker-based systems?

Independent variables	Dependent variables
Monitoring tool • Elastic stack • Netdata • Prometheus • Jaeger • Zipkin • baseline	Energy efficiency
	CPU usage
	CPU load average
	RAM usage
	Network traffic
	Execution time
Frequency - the scrape, or sampling interval	
Workload - the number of virtual users	

Subject
Train Ticket System
• Medium-size microservice benchmark system for booking train tickets
• Previous usage in empirical studies • Open-source • Representative through size (41 microservices), granularity and variety of microservices

Example (ICSOC 2024)



Q2

I am having trouble to decide whether my other factor other than the main one is a co-factor or a blocking factor. Could you please explain again how we should decide whether a factor should be a co-factor or a blocking factor?

Answer

Co-factor:

- you are interested in assessing how it interacts with your main factor
- it is part of the research questions
- your experiment is not a 2F anymore → you start thinking about ANOVA

Blocking factor:

- you suspect it might have an influence on your dependent variablea, but you you use them only for “compartmentalizing” your experiment
- you apply statistical the same statistical analysis inside each block

Q3

More examples of hypothesis testing (especially where the null hypothesis gets rejected)

ICSOC 2024

6.1.5 Statistical hypotheses. To answer **RQ1** concerning **energy efficiency**, we formulate the following hypotheses, where μ_t is the mean energy efficiency (i.e., total energy consumed by the TTS, during a load test), when monitoring the system with tool t:

- Null hypothesis: Energy efficiency does not significantly differ among the benchmark system being monitored using different monitoring tools.

$$H_0^{energy} : \mu_i = \mu_j, \quad \forall\, i, j \in \{baseline, Elasticsearch, Netdata, Prometheus, Jaeger, Zipkin\}, i \neq j \tag{2}$$

- Alternative hypothesis: Energy efficiency significantly differs for at least one pair of monitoring tools.

$$H_a^{energy} : \exists\, \mu_i \neq \mu_j, \quad where\ i, j \in \{baseline, Elasticsearch, Netdata, Prometheus, Jaeger, Zipkin\}, i \neq j \tag{3}$$

Q3

More examples of hypothesis testing (especially where the null hypothesis gets rejected)

ICSOC 2024

Table 3. Results of Shapiro-Wilk (SW) and Kruskal-Wallis (KW) tests for each frequency (F) and workload (W) block. Bold text denotes a significant difference ($\alpha = .05$)

Block	SW (p-value)	KW (p-value)	η^2	η^2 interpretation
F Low, W Low	**0.00113**	**0.00156**	0.3	large
F Low, W Medium	0.0939	0.21	0.0413	small
F Low, W High	**0.0157**	**3.77e-06**	0.59	large
F Medium, W Low	**0.0172**	**0.00157**	0.299	large
F Medium, W Medium	**0.019**	0.303	0.0189	small
F Medium, W High	**0.00228**	**1.17e-06**	0.645	large
F High, W Low	**9.25e-05**	0.154	0.0594	small
F High, W Medium	0.0826	**0.022**	0.165	large
F High, W High	**9.52e-05**	**4.58e-07**	0.69	large

Q4

I would appreciate knowing how to classify code as "memory-bound" or "compute-bound" from Prof. Ivano's perspective, and how to support this classification with concrete evidence.

Answer

Given a certain task, the answer comes from empirically measuring it (e.g., via `top`):

- **Memory-bound** = when the memory [bus] is congested by limiting the rate of execution of compute instructions
- **Compute-bound** = primarily when the ALUs (arithmetic logic units) of the processor are fully utilized and unable to provide additional throughput

Konstantinidis, Elias, and Yiannis Cotronis. "A practical performance model for compute and memory bound GPU kernels." 2015 23rd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing. IEEE, 2015.

Q5

I would like to see a practical example of how should one move from the experiment phase into analyses phase. Guidings and tips on what to do with that raw that that we would get after the experiment, and how to properly analyze it afterwards. A holistic view of the whole experiment

Answer

If you configure Experiment Runner properly, your run table will be already complete and ready to be imported into an R dataframe.

Tips:

- check if your populated run table contains all the rows and columns you expect for running your statistical analysis
- always search and analyze outliers (column by column)
  - decide what to do with rows with outliers
- check also min and max values for each block and dependent variables
- visualize visualize visualize!

Q6

Can you explain a bit on MF-MT and the types of analysis we can perform like Anova or Kruskal-Wallis and how we can understand which one to do?

Answer

See next slides for a refresher

Examples of studies using ARTool:

- EASE 2023
- ICT4S 2024

>=1 factors - >2 treatments

Design	Parametric	Non-parametric
One factor, one treatment		Chi-2, Binomial test
One factor, two treatments, completely randomized design	t-test, F-test	Mann-Whitney, Chi-2
One factor, two treatments, paired comparison	Paired t-test	Wilcoxon, Sign test
One factor, more than two treatments	ANOVA	Kruskal-Wallis, Chi-2
More than one factor	ANOVA[a]	

ANOVA (ANalysis Of VAriance)

Goal: understand how much of the total variance is due to differences within factors, and how much is due to differences across factors

- Many types of ANOVA tests
- Works for many experiment designs

Hypotheses:

H_0: $\mu_1 = \mu_2 = \mu_3$ H_a: $\mu_1 \neq \mu_2 \vee \mu_1 \neq \mu_3 \vee \mu_2 \neq \mu_3$

Parametric

Significance

F tends to be larger if H_0 is false

→ the more F deviates from 1, the stronger the evidence for unequal population variances

- Methods to determine significance level:
  - *textbook*: compare F against a table of critical values (according to DF and α). If $F > F_{critical}$, reject H_0
  - ***computer-based:*** compute the p-value of finding F greater than the observed value. If $p < \alpha$, reject H_0

Types of ANOVA

- *One-way* ANOVA
  - one factor, >2 treatments
  - if 2 treatments: equivalent to *t-test* (almost never used)

```
> summary(data$Watts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  207.3   214.0   214.2   215.7   219.8   222.2
> summary(data$Case)
mysql_modified mysql_original  mysql_vanilla
            10             10             10
```

```
#one-way
data <- read.csv('practice_1_power.csv')
data.aov <- aov(Watts~Case, data=data)
summary(data.aov)
```

```
> summary(data.aov)
             Df Sum Sq Mean Sq F value   Pr(>F)
Case          2  232.9   116.5   14.38 5.59e-05 ***
Residuals    27  218.6     8.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Types of ANOVA

- *Factorial* ANOVA
  - 2 (two-way) or more factors
  - any number of treatments
  - also computes interactions

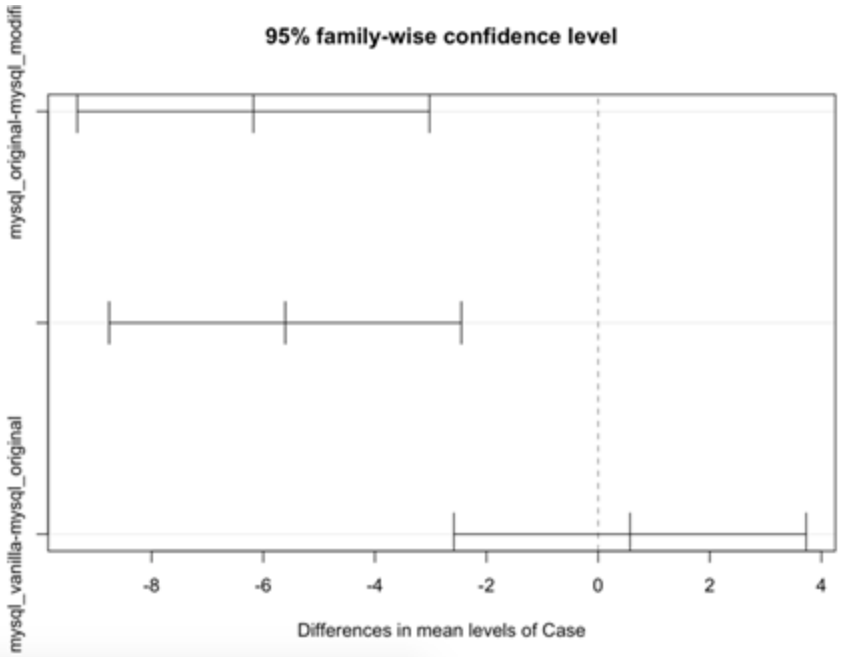
```
# #two-way
server <- factor(sample(1:3, 30, replace=TRUE), levels=c(1:3), labels=c('Server 1', 'Server 2', 'Server 3'))
data_new <- cbind(server, data)
data.2aov <- aov(Watts~Case*server, data=data_new)
summary(data.2aov)
#
```

```
> summary(data.2aov)
             Df Sum Sq Mean Sq F value   Pr(>F)
Case          2 232.95  116.48  15.421 7.57e-05 ***
server        2  32.22   16.11   2.133    0.143
Case:server   4  27.80    6.95   0.920    0.471
Residuals    21 158.62    7.55
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How to know which treatments really differ?

Tukey's test

```
summary(data.aov)
posthoc <- TukeyHSD(x=data.aov, 'Case', conf.level=0.95)
plot(posthoc)
```




```
> posthoc
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Watts ~ Case, data = data)

$Case
                                diff       lwr       upr     p adj
mysql_original-mysql_modified -6.176 -9.331299 -3.020701 0.0001302
mysql_vanilla-mysql_modified  -5.605 -8.760299 -2.449701 0.0004301
mysql_vanilla-mysql_original   0.571 -2.584299  3.726299 0.8953945
```

ANOVA assumptions

- The dependent variable should be **continuous**
- Samples must be **independent**
- **Normal distribution** of the dependent variable between the groups (approximately)
- Residuals (aka errors in the sample) should be normally distributed
  - qqPlot(residuals(myData.aov))
- **Homoscedasticity**
  - variance between groups should be the same
    - leveneTest(x ~ y, data=myData)

Assumptions violated

→ non-parametric alternative

ANOVA: non-parametric alternative

- Kruskal-Wallis: one-way non-parametric ANOVA
  - one factor, multiple treatments
  - no estimate of the treatment effect (due to ranking)

```
#non-parametric one-way
kruskal.test(Watts~Case, data=data)
```

```
> kruskal.test(Watts~Case, data=data)

        Kruskal-Wallis rank sum test

data:  Watts by Case
Kruskal-Wallis chi-squared = 12.718, df = 2, p-value = 0.001732
```

Use ARTool when you have >2 factors

Non-parametric

TODO

Do you feel that certain metrics (like CPU Usage , Memory Usage) have more influence on the calculation of energy efficiency over other metrics (like Lines of Code) for our experiment? If so, how can we determine which metrics have more influence and define how some have more impact?

Answer

Typically (but this is not the rule):

- CPU usage and execution time correlate strongly with energy
  - Think about the `E = P*t` formula
- Memory usage tends to do not correlate with energy
- Lines of code is completely orthogonal and it depends on what the program actually does, the used programming language, compiler optimizations, etc.

TODO

Requirements regarding the data analysis part in our project.
How deep do we need to analyze?

Answer

In general, we expect at least:

- data exploration with summary statistics + plots
- hypotheses testing (with checks on assumptions of statistical tests + transformations
- effect size estimation
- elaboration on the possible causes of the obtained statistical results

Check our latest studies in the “Articles on performed experiments” folder in Canvas

TODO

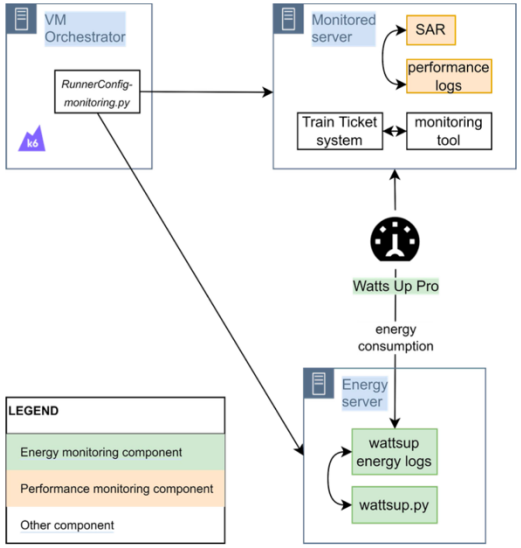
There seems to be some confusion regarding the data collection process (using a Raspberry pi vs our laptops) for energy efficiency experiments.

ICSOC 2024

RQ3 Experiment Execution

RQ3: What is the runtime overhead of monitoring tools for Docker-based systems?

- Dedicated measurement infrastructure in the Green Lab research laboratory
- Experiment Runner for automatic execution of the experiment
- K6 for load testing the Train Ticket System
- Watts Up Pro power meter to measure energy at the machine level
- SAR Linux system utility to measure performance



Measurement infrastructure