Final Project

Failure Data and Reliability Analysis for Software-as-a-Service (SaaS) Business Application

ECE 313

Probability with Engineering Applications

Group Zeta

Daniel Petrisko

Meruyert Mussakhanova

Dept. of Electrical and Computer Engineering

University of Illinois at Urbana Champaign

Data collection technique

- Logs from (SaaS) Business Application
- Time period: 10 months, from 1 January, 2012 to 1 November, 2012
- Source: Provided by Dr. Catello Di Martino

Data description

- 2 log files: data_size contains file size information, data_failure - work status: “Failed”/ “Succeeded”
- Failure/Success status has been derived from data_size file for further analysis

fileID	event_start_time	event_end_time	status_code	failure_reason
875490	"2012-01-01 00:00:26"	"2012-01-01 00:00:26"	SUCCEEDED	
875491	"2012-01-01 00:00:26"	"2012-01-01 00:00:55"	SUCCEEDED	
875492	"2012-01-01 00:00:26"	"2012-01-01 00:00:26"	SUCCEEDED	
875493	"2012-01-01 00:00:26"	"2012-01-01 00:00:26"	SUCCEEDED	
875494	"2012-01-01 00:00:55"	"2012-01-01 00:00:55"	SUCCEEDED	
875495	"2012-01-01 00:00:55"	"2012-01-01 00:00:58"	SUCCEEDED	

data_failure file example

file_event_id	file_event_start_time	file_event_end_time	file_row_count	status	load_row_count	file_size_kb
307598	"2012-01-01 00:44:27"	"2012-01-01 00:45:12"	420	PRC	0	49
307635	"2012-01-01 01:01:13"	"2012-01-01 01:03:10"	8480	PRC	8480	819.274
307639	"2012-01-01 01:04:11"	"2012-01-01 01:10:02"	349533	PRC	349533	35590.1
307641	"2012-01-01 01:04:13"	"2012-01-01 01:08:12"	9932	PRC	0	418
307642	"2012-01-01 01:04:14"	"2012-01-01 01:09:07"	21545	PRC	21545	2248.78

data_size file example

Goal and Analysis approach

Goal: derive the hazard rate and reliability of the system, and predict the failure rate depending on Workload Volume and Workload Intensity.

- Choose two random variables: Workload Intensity = number of files per hour and Workload Volume = size of files per hour, and perform analysis based on them
- Split the data into 1-month intervals for faster processing and convenient results representation.

Concepts from class and tools used

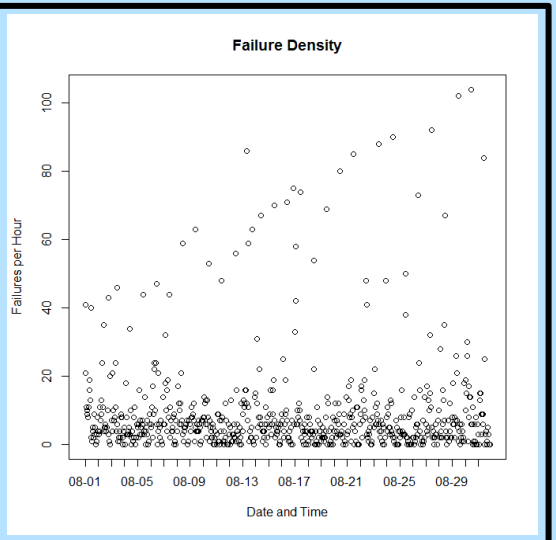
- Failure Frequency
- Discrete Random Variable
- Joint Distribution
- Hazard rate
- Failure Density
- Reliability

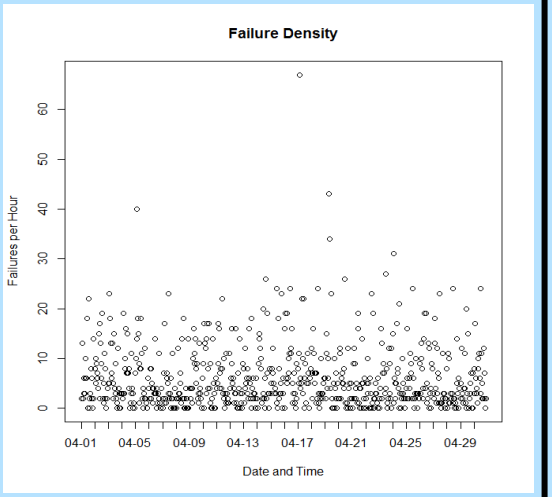
- Tools: R for statistical programming. Reason: has capability to operate on large datasets.

Results of analysis

1. Failure frequency analysis

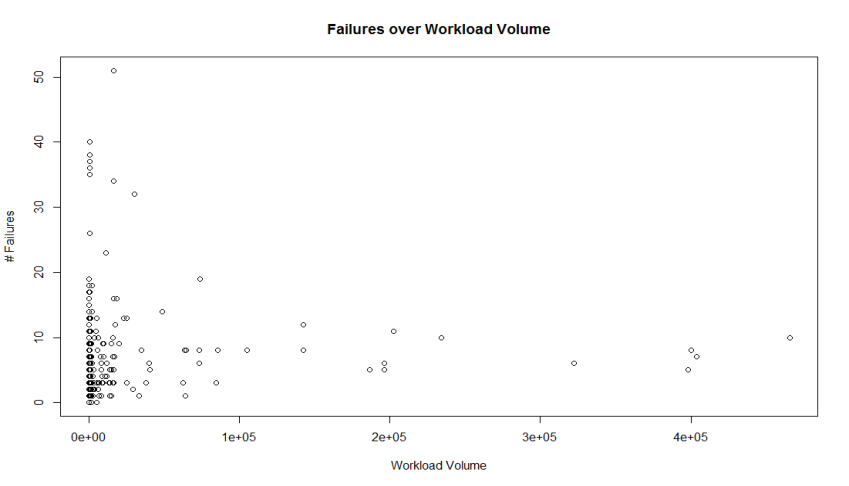
- Failures are uniformly distributed over each month, with little variation.
- Some months, such as February and August have steadily increasing outliers

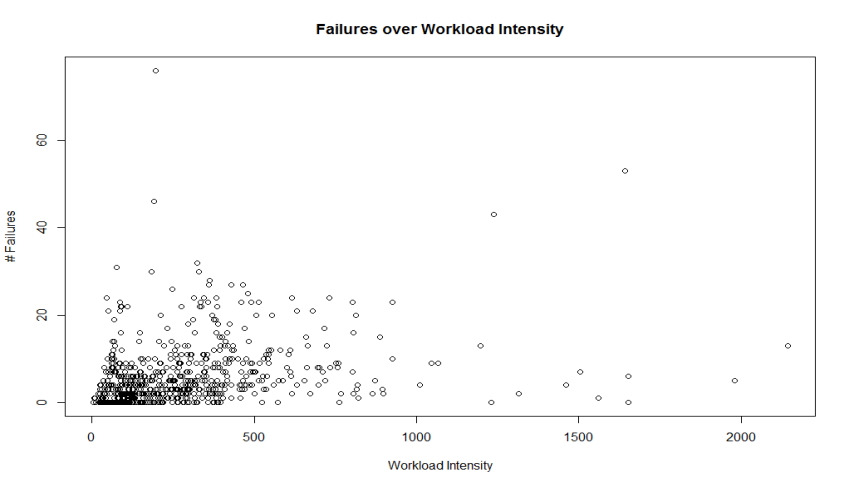




2) Failure distribution vs Workload Intensity and vs Workload Volume

- In general Failures over Workload Intensity are distributed more evenly than Failures over Workload Volume
- More failures when the size of sent files is small (Workload Volume)

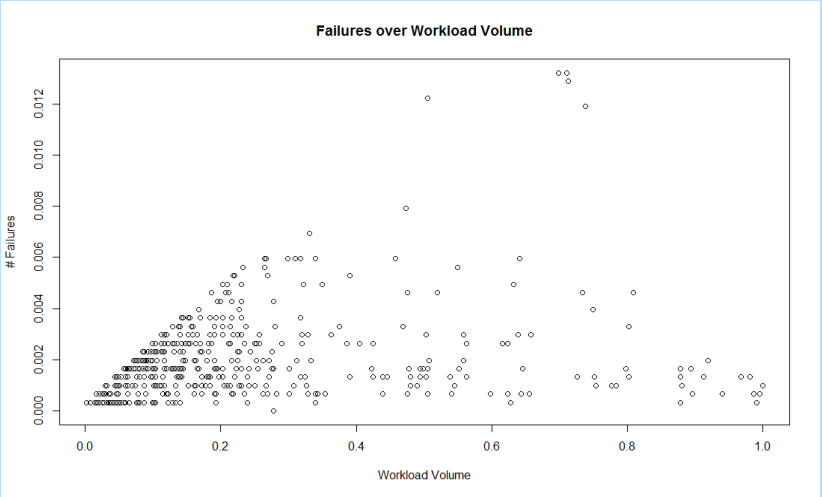


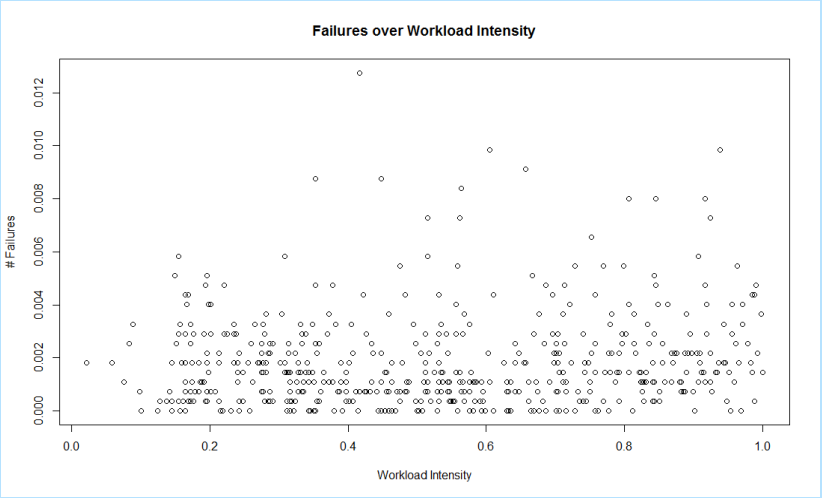


3) Joint distribution of failures based on random variables Workload Intensity and Workload Volume

1. Normalized Failure Density over Workload Intensity and Workload Volume

- Failure Density is proportional to Workload Volume
- Failure Density seems to be unrelated to Workload Intensity

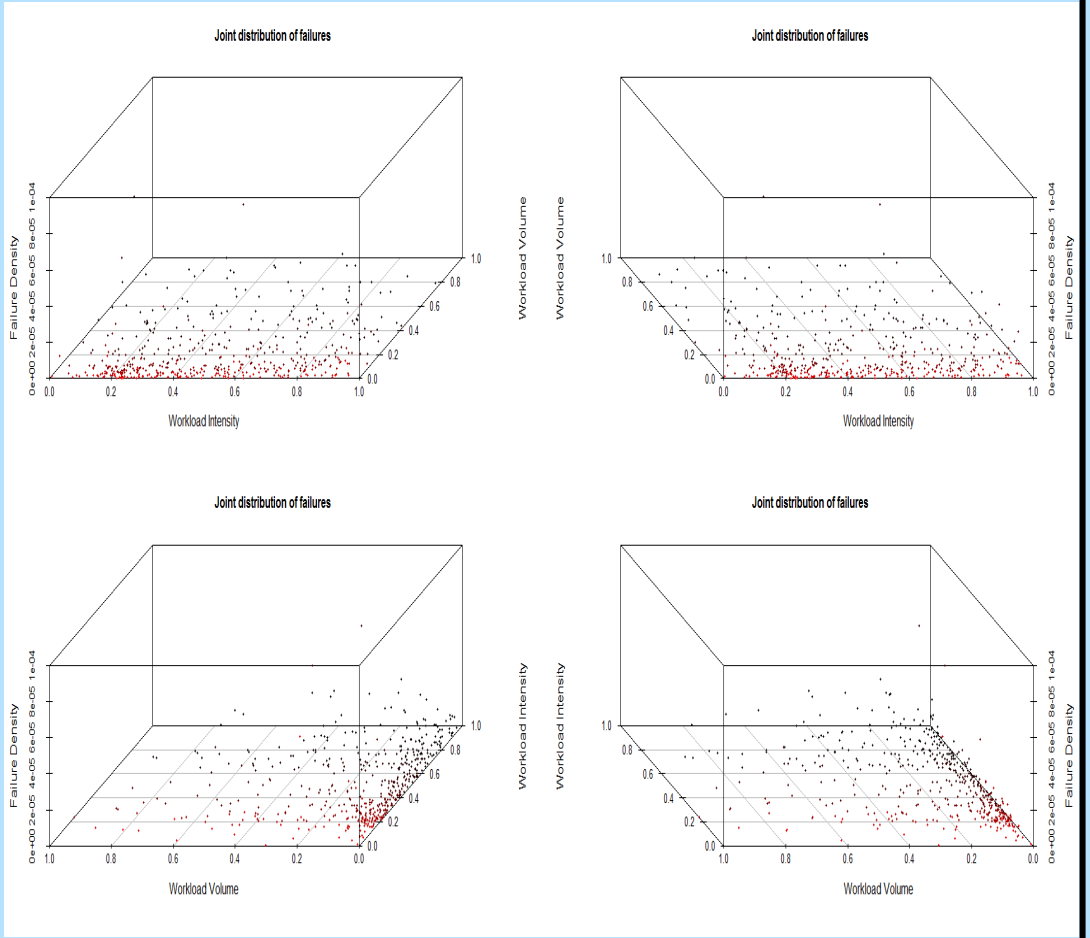




3) Joint distribution of failures based on random variables Workload Intensity and Workload Volume

2. Build joint distribution matrix
3. Plot joint distribution

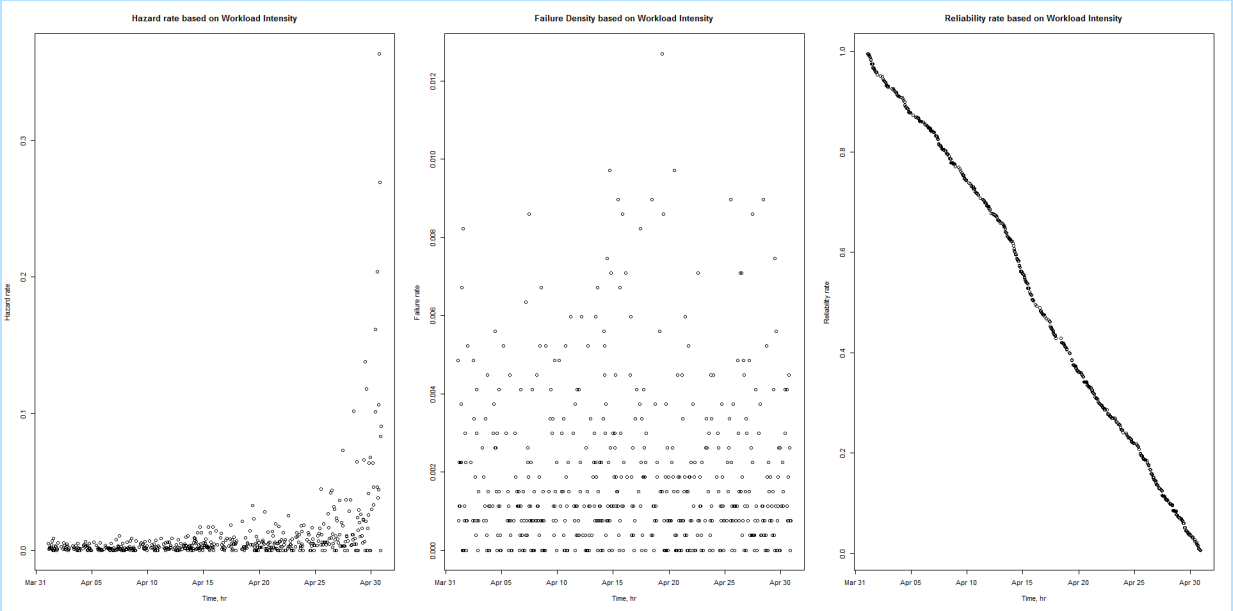
Generated plots confirm that the Failure Density is higher when the size of files sent is small Failure Density does not depend on number of files sent.



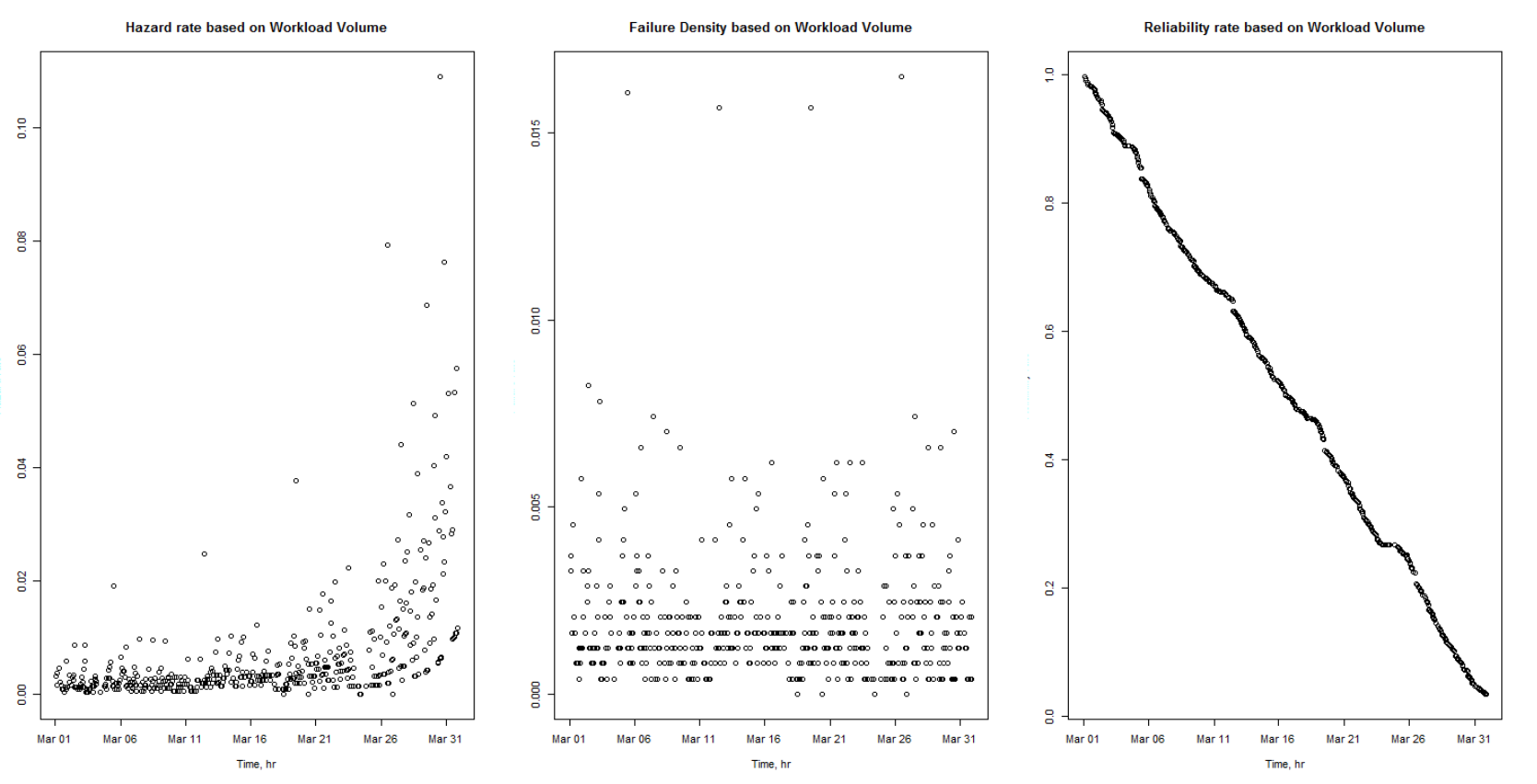
4) Hazard function, failure rate and reliability

Notable Features

- Exponentially increasing Hazard Rate z(t)
- Uniformly distributed Failure Density Function fd(t)
- Reliability Function R(t) nearly linear decreasing but is piecewise exponential for certain days
- This indicates daily wear on the system causing lower reliability



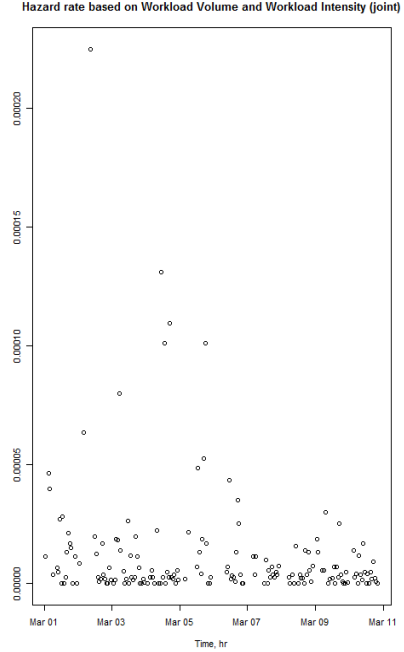
4) Hazard function, failure rate and reliability (continue)

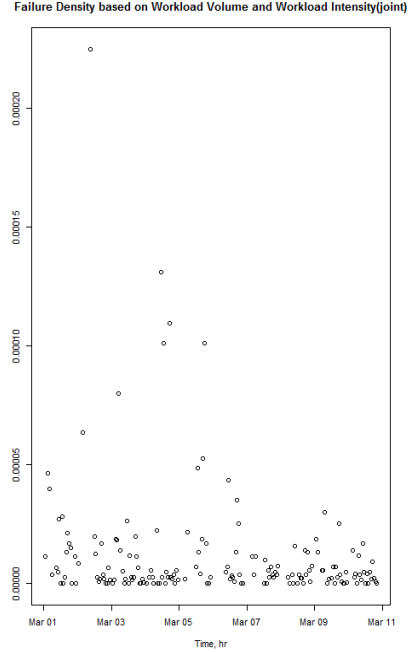


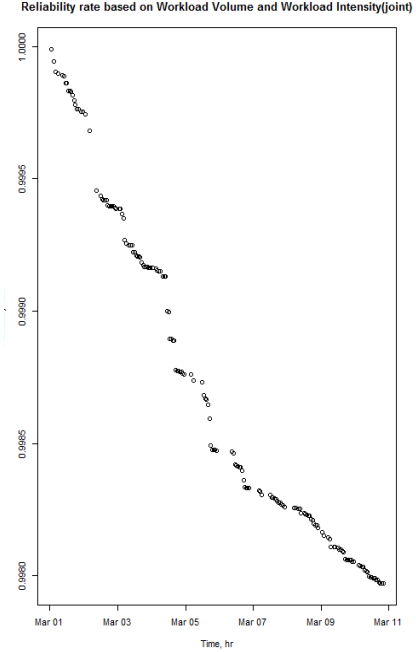
Hazard function, failure rate and reliability (continue)

Notable Features

- Uniformly distributed Hazard Rate z(t) and Failure Density Function fd(t)
- Reliability Function R(t) is close to exponential function







Summary and Key insights:

We accomplished the goal of the project, which was to derive the hazard rate and reliability of the system, and ability to predict the failure rate depending on Workload Volume and Workload Intensity.

We conclude that the failure rate has high correlation with the Workload Volume, but almost no relation to the Workload Intensity.

However, when we normalized Workload Intensity and Workload Volume and examined the Failure Rate with respect to this joint distribution (the 3d plots), it made correspondence between Failures and Volume more clear.

The Reliability Function of the system shows constantly decreasing reliability, with periodic exponential streaks, possibly indicating daily wear on the system.

Suggestions:

- Give more guidance on selecting data sets, so there is less emphasis on data collection
- Start the final project earlier, so as to space out presentations and progress reports and give more time for analysis
- Give feedback on performance after each project progress presentation