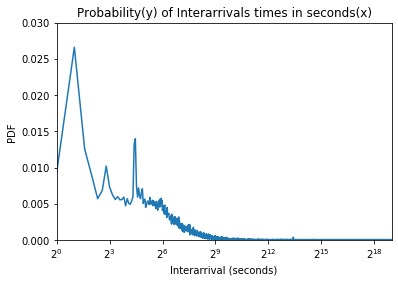
December 23

University of Haifa

Workload for Computer Systems Performance Evaluation

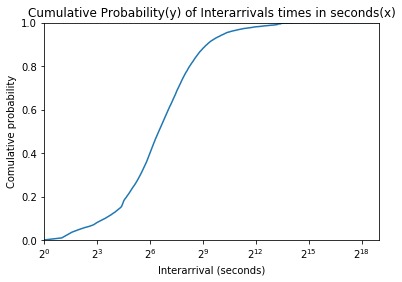


**NASA-Log File Graphs:**

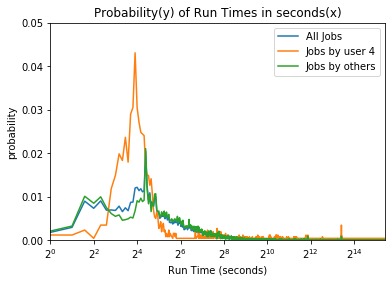


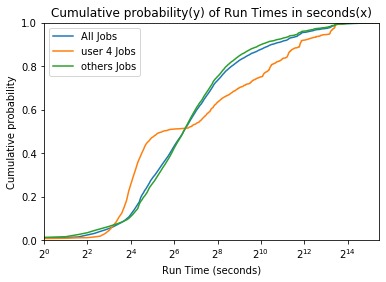
**(1)**

אנחנו יכולים לראות שהגרף מיוצג בצורה הלוגריתמית בציר ה-x והסיבה היא שהיו ערכים ממש גדולים שאי אפשר לייצג עם הערכים הרגילים, בנוסף שמנו לב שהיה משתמש בעל שם משתמש “special” שהיו לו JOBS ארוכים במיוחד ואחד מהם נמשך כמעט ל 6 ימים. זה בעצם מסביר את ה Heavy tail שאפשר לראות בגרף.



**(2)**



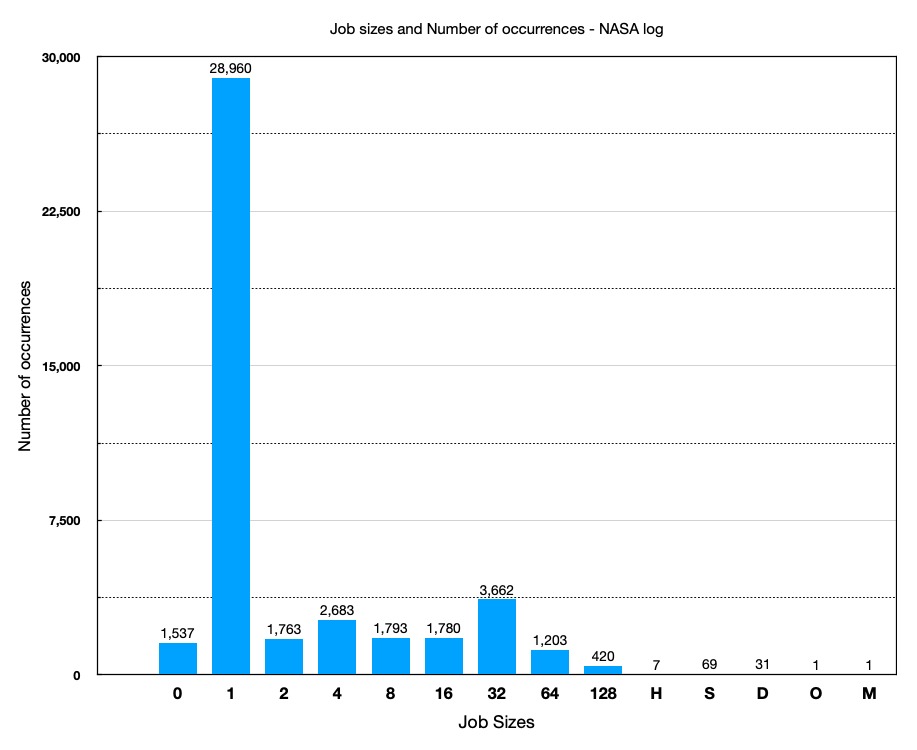


**(3)**

**(4)**

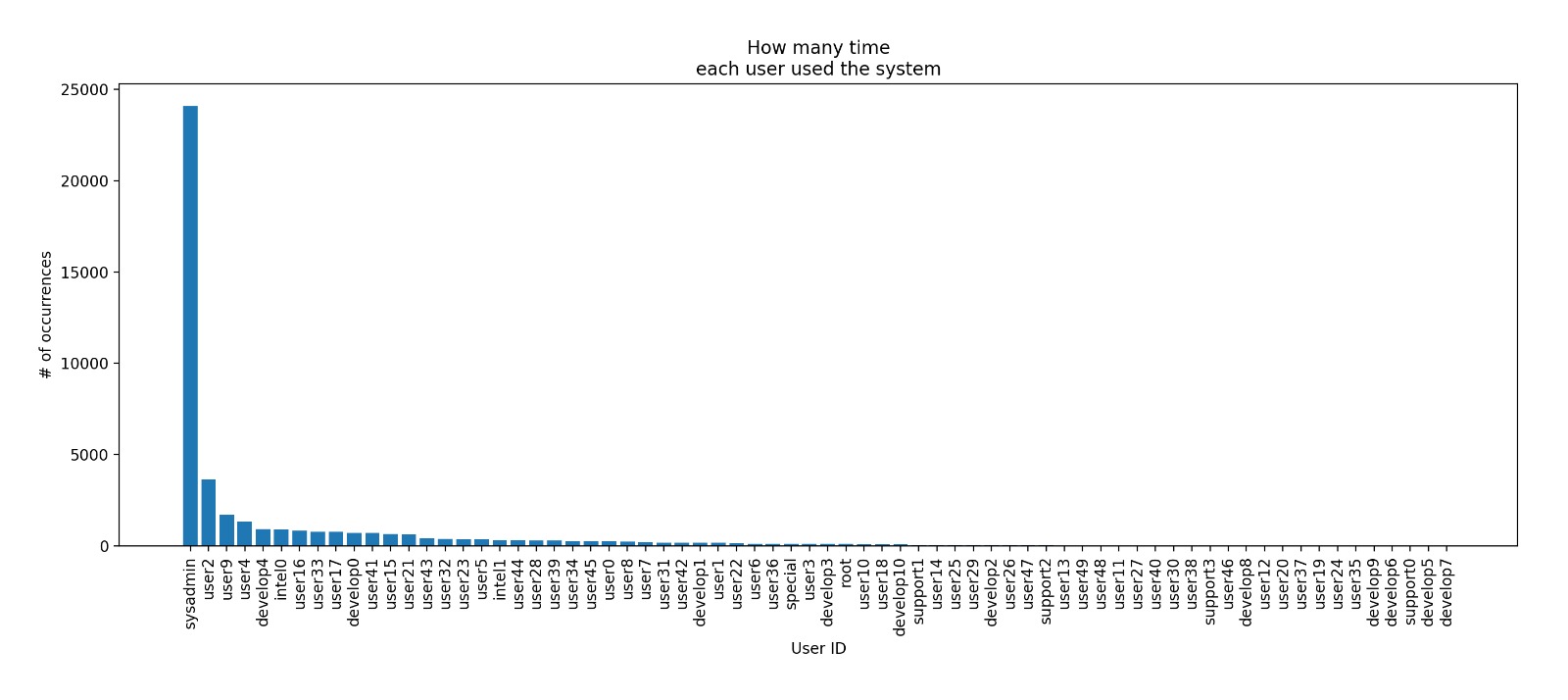
גרפים שמראים התנהגות שונה של אחד הusers משאר הusers שזמני הריצה של הjobs שלו יותר גדולים משל השאר.

אפשר לראות גם שבזנב של הקו של משתמש “special” שאפילו כאשר את ה Y עדיין לא שווה ל 1 וזה יוצא מהסיבה שיש לא JOB שנמשך לכמעט 6 ימים.

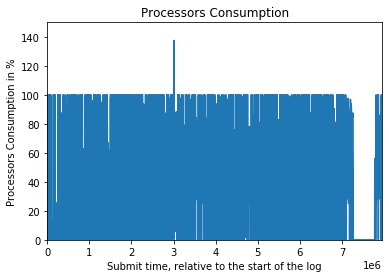


**(5)**

היסטוגרמה שמראה את מספר המופעים של גדלי job שונים. אפשר לראות שהגודל הכי נפוץ הוא 1 וזה נובע מכך שיש המון פעולות של sysadmin (אפשר לראות אותם בגרף הבא) שהם פעולות מערכת, אז בדרך כלל יהיו פעולות קצרות כמו למשל pid שבודקת פעילות המערכת.



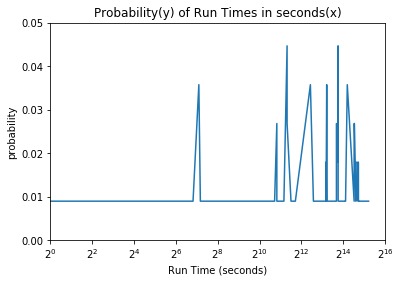
**(6)**



**(7)**

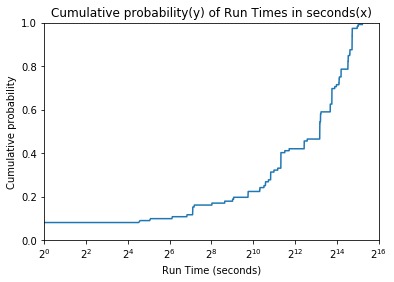
הגרף הזה מראה את אחוז המעבדים התפוסים בכל שניה מהתחלת ה Log ועד סיומו.

(המספרים בציר ה X הם כפולות של )

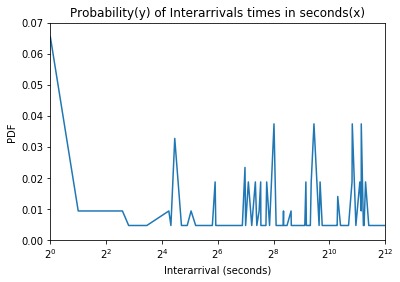
**MATLAB-Log File Graphs:**

**(8)**

אפשר לראות בגרף ה PDF כל מיני bursts ו flurries שבעצם מראות שיש כמה RunTimes נפוצות.

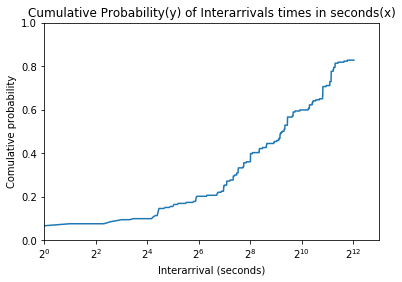


**(9)**

אפשר לראות בגרף ה PDF כל מיני bursts ו flurries שבעצם מראות שיש כמה Interarrival Times נפוצות.

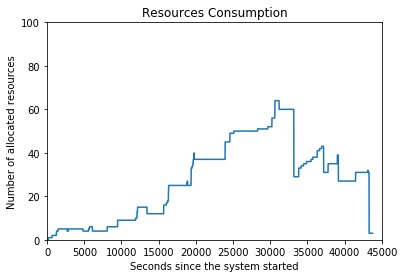
**(10)**

אפשר גם לראות שכאשר X=1 יש לנו הסתברות גדולה יחסית לשאר ה-Interarrivals שזה בעצם אומר לנו שיש כל מנה jobs שבאים אחד אחרי השני בשניה אחת.



**(11)**

**(12)**

ניתן להסיק מהגרף ששעות העומס על המערכת הן בין 17:40-20:05 וזה הגיוני שהעומס ירד באופן משמעותי אחרי השעה 20:00 כי בדרך כלל חלק גדול מהמשתמשים (עובדי היי-טק וסטודנטים) מסיימים עבודה בסביבות השעה הזו.

או ייתכן שהשרת של הMATLAB שחרר כל מיני משאבים ביחד באותו זמן.

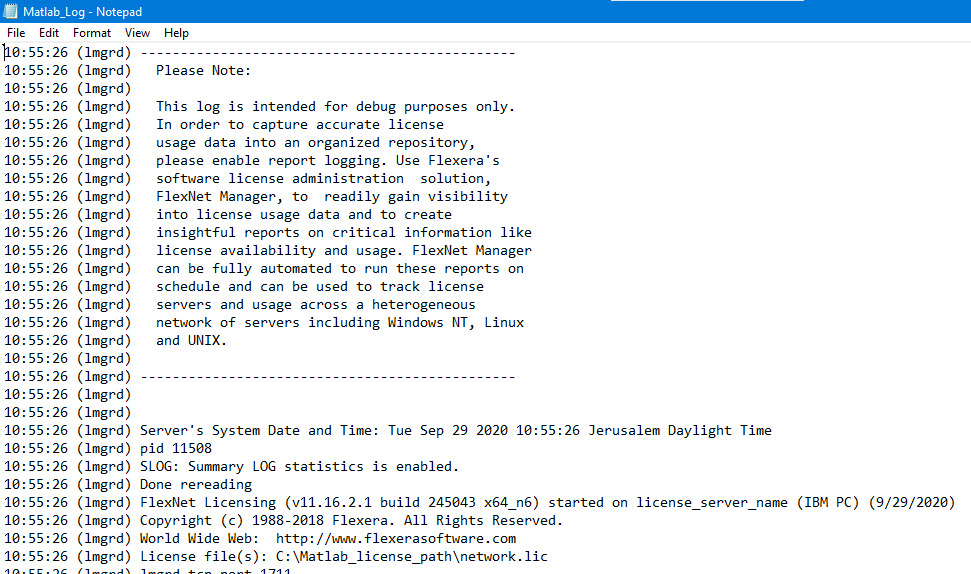
**Stage 2 – Questions:**

שאלה 2:

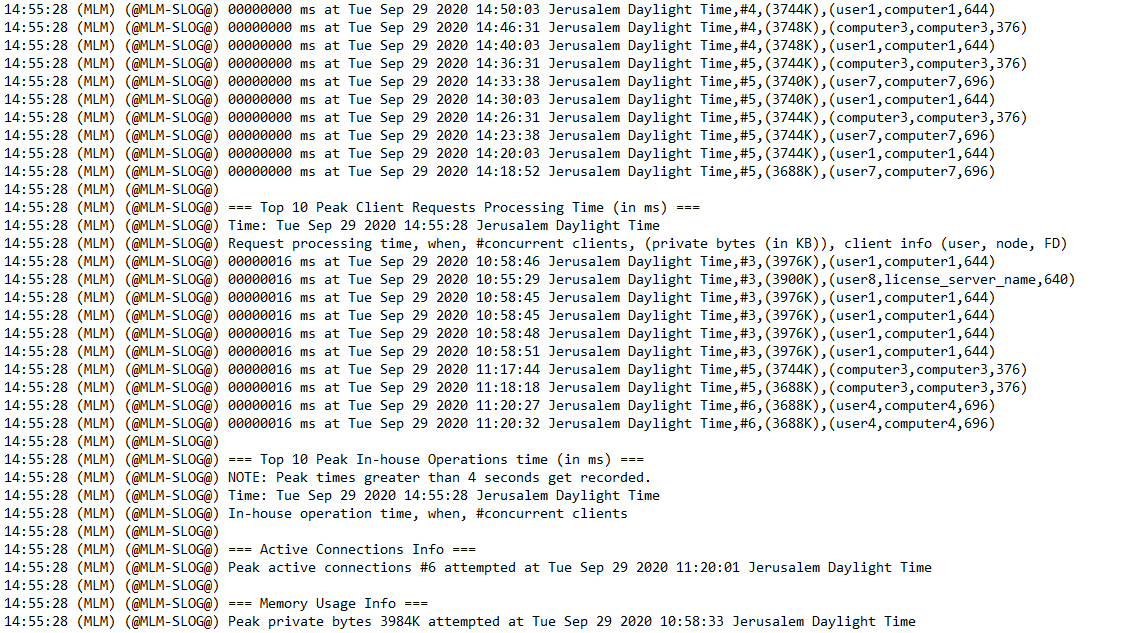
נקח למשל את הגרף של הProcessors Consmuption של הלוג של NASA- ניתן לראות כי יש לנו פיק של 140% של Processors שזאת תופעת Abnormal Behaviour. בגלל הסיבה הפשוטה שבכל רגע נתון, הגרף הנ"ל מייצג את מספר המעבדים ה"תפוסים" ע"י הג'ובים השונים, לא ניתן ל"שים את האצבע" על הבעיה והג'וב הספיציפי שגורם לפיק הזה. (בספר בעמוד מספר 50- פיסקה 2).

שאלה 3:

בטרייס של NASA, קיבלנו אותו נקי ומוכן לשימוש, לכן לא הצטרכנו לעשות הרבה בנדון, אלא כבר ליישם את הנדרש (SWF, גרפים וכו').

בטרייס של הMATLAB שקיבלנו, אכן היה לנו כל מיני שורות שהיינו צריכים לנקות – כגון:

תיעוד של השרת (התחלה) ושל הTOOLBOXES הזמינים שנטענו.



סיכום מעת לעת של המשתמשים ושל החיבורים הפעילים במערכת כרגע.

כמובן שלאחר הניקוי, היה לנו מידע נטו שנוכל להשתמש בו, לצורך בניית ה- SWF, כי כל שורה שהיא מהצורה של IN/OUT השארנו, כך שנוכל לקבל את מלוא המידע הדרוש במשימה הזאת ולמשימות הבאות.

אין כל כך מה להשוות בין התוצאות שהיינו מקבלים בלי הניקוי לעיל, כי כל מה שניקינו – יכולים לקבל מהמידע שנשאר (חיבורים פעילים, שמות משתמשים, שעת התחלה סיום וכו'...).

לכן, התוצאה הסופית של הלוג הנקי הייתה רק שורות של IN OUT כך שנוכל לדעת עבור כל משתמש וכל משאב מי, מתי וכמה השתמשו בו למשל.

**Stage 3:**

אפשר לראות שהגרפים ממוספרים מ 1 ל 12 נתחיל לתאם כל גרף להתפלגות מסוימת.

**(1):Multiple Erlang distributions**

ניתן לראות שגרף זה "מורכב" משתי התפלגויות Erlang, כאשר הפרמטר יהיה .

**(2) :CDF of Erlang distribution**

באותו אופן לגרף מספר 1, נראה כי גרף זה מייצג לנו CDF של התפלגות Erlang כאשר הפרמטר יהיה .

**(3) with tail of exponential distribution : More than one Hyper-Erlang**

ניתן לראות שגרף זה הינו שילוב של שתי התפלגויות – התפלגות Erlang של user רגיל והתפלגות Erlang של special user. לכן, שילוב של שתיהן ייתן לנו Hyper-Erlang distribution, כאשר הפרמטרים יהיו

**(4): CDF of Hyper-Erlang**

באותו אופן לגרף מספר 3, נראה כי גרף זה מייצג לנו CDF של התפלגות Hyper-Erlang כאשר ערכי הפרמטרים זהים לגרף ה-PDF.

**(5) אין התפלגות שמתאימה להיסטוגרמות**

**(6) אין התפלגות שמתאימה להיסטוגרמות**

**(7) אין התפלגות שמתאימה להיסטוגרמות**

**(8) lognormal מעורב:**

ניתן לראות שגרף זה מתפלג בצורה "יחסית" נורמלית, אך גרף זה נמצא במרחב לוגריתמי (ציר ה-x הינו סקאלה לוגריתמית), ולכן גרף זה מתאים להתפלגות לוג-נורמלית. בנוסף נשים לב שה-mean וה- standard deviationגם נמצאים במרחב לוגריתמי.

(9) -------------

לפי המידע על התפלגות הלוג-נורמלית, אין צורה חד-משמעית לגבי ה-CDF.

(10) -------------

גרף זה מטעה קלות – ראשית, נשים לב שהוא נמצא במרחב לוגריתמי. אם נסתכל על ה-interval שנמצא מ- , ניתן לחשוב שמדובר בהתפלגות לוג-נורמלית (מרחב לוגריתמי, התפלגות "יחסית" נורמלית).  
מצד שני, ה-interval שנמצא בתחילת הגרף עד לא ברור לגמרי.

(11) -------------

**(12) Normal distribution:**

ניתן לזהות בגרף זה "עקומת פעמון" אשר מאפיינת את העומס על המערכת (הסקנו שהעומס נובע ישירות מפיזור שעות העבודה בכל יום – עלייה בבקרים וירידה בערבים).

**Stage 4:**

**NASA-Log File Graphs**

**A picture containing chart

Description automatically generated*CDF of Runtimes time:***

**Engineering drawing

Description automatically generatedAll Users**

**Shape

Description automatically generated**

**Chart, line chart

Description automatically generatedChart

Description automatically generatedOther Users**

**Shape, polygon

Description automatically generated**

**A picture containing line chart

Description automatically generatedA picture containing chart

Description automatically generatedSpecial User**

**Shape, arrow, polygon

Description automatically generated**

***CDF of Interarrival time*Chart, histogram

Description automatically generatedShape

Description automatically generated:**

**Chart

Description automatically generated**

**MATLAB-Log File Graphs**

***Chart

Description automatically generatedCDF of Runtimes:Chart

Description automatically generatedChart

Description automatically generated***

***Chart, scatter chart

Description automatically generatedCDF of Interarrival time:Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated***

**Stage 5:**

נראה שגרף מספר 6 (כמות jobs לכל user במערכת) יכול להיות מתואר לפי התפלגות Zipf.

בדומה לגרף Zipf של הסתברות הופעות מילים בקטע טקסט, נייצר גרף לפי אותו אופן. באמצעות גרף מספר 6 אשר מתאר לנו את כמות ה-jobs לכל user, נוכל למצוא את ההסתברות להופעת job של user ספציפי.

A picture containing application

Description automatically generatedנציג זאת בגרף:

Chart, line chart

Description automatically generatedציר ה-x יהיה המספר הסידורי של ה-users, וציר ה-y יהיה ההסתברות להופעת job מאותו user. הקו האדום מתאר את הפונקציה שנוצרת משימוש בהתפלגות Zipf על הערכים הנ"ל.  
  
בשביל לראות יותר בנוחות את הגרף, נהפוך את העמודות לקו רציף ונייצג את צירים בסקאלה לוגריתמית:

ניתן לראות שההתאמה לא טובה כפי שציפינו. ננסה לקבץ מספר דגימות יחדיו בשביל התאמה טובה יותר.

Chart, line chart

Description automatically generatedבגרף הבא, נעשה קיבוץ על 5 דגימות. ניתן לראות שישנו שיפור קל.

Chart, line chart

Description automatically generatedננסה לקבץ כמות גדולה יותר של דגימות – הפעם נקבץ 7 דגימות ונראה את הגרף הבא:

נשים לב שה-fitting בגרף עם קיבוץ של 7 דגימות יותר מתאים לפונקציית ה-Zipf.

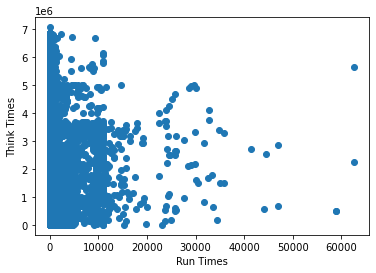
**Stage 6:**

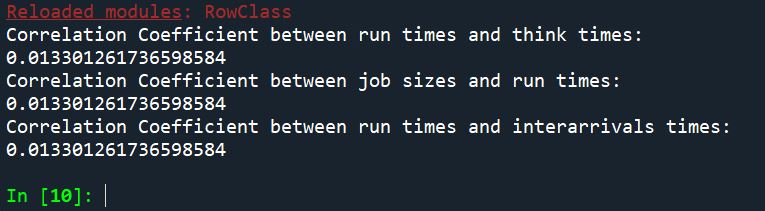
**NASA**

If we look at the CDF graph of the run times, we can see that most of the jobs have run time less than 12,000 (less or more).

On the other hand, users in NASA don't like to wait until the previous job is completed and they start to run in parallel with the previous jobs, as a result, we get that its not a must that longer jobs have longer interarrival time.

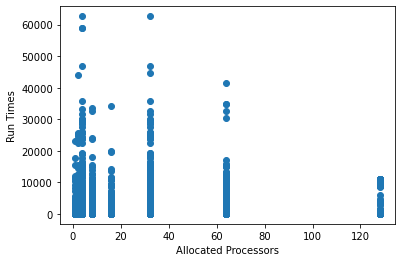
This can be also noticed in the CDF graph of interarrival times.





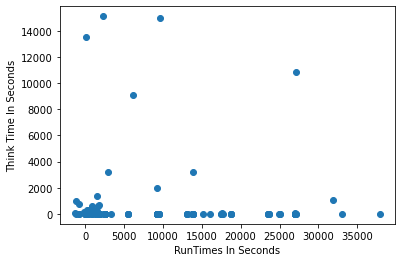
Chart, scatter chart

Description automatically generated

****

**Matlab**

**Run Times – Think Times Correlation Coefficient:** -0.06555923348325185



**Chart, line chart

Description automatically generatedStage 7:**

**Chart, histogram

Description automatically generated**