CSC498 - Simulation Report

Daniel Gabriele | Justin Pham | Roy Lu

1. Harchol-Balter, M. (2014). *Performance modeling and design of computer systems: queueing theory in action*. New York: Cambridge University Press.

Table of Contents

1. Harchol-Balter, M. (2014). *Performance modeling and design of computer systems: queueing theory in action*. New York: Cambridge University Press.

I – *Abstract*

Rising incidents of delayed shipments from the manufacturing of commercial and military aerospace related goods is an increasing concern in international government departments relating to travel, commerce, and defense. The purpose of this report is to investigate the connection between the utilizations of the different departments responsible for the production of landing gear at the Collins Aerospace Oakville plant and if any possible bottlenecks in their manufacturing and rework process can be identified. Using a classed Jackson-network analysis, this study analyzed the time between manufacturing events as well as rework events while noting how much time each department spent to start, process, and finish each event. From the simulation, the departments with the highest utilization were "Site Quality" and "Manufacturing" and are most suspect of being the bottleneck in the production process. An interesting trend seen is that the liaisons, who are MRB QAs, between Site Quality, Manufacturing, and Engineering had the single-highest utilization than any other position at the plant with a utilization of 93%. While this report successfully identifies which departments are taking the longest during the rework process, further research must be conducted to identify why there is an extremely large number of defects and, by extension, reworks appearing in the first place.

II – *The Manufacturing Process*

A Jackson-network analysis was used for this problem considering that, much like many other systems in our society, Collins Aerospace uses a queue system for all of customer orders on

1. Harchol-Balter, M. (2014). *Performance modeling and design of computer systems: queueing theory in action*. New York: Cambridge University Press.

site. For each part type labelled A, B, C, D there is a main manufacturing process that is unique to each part type that must be adhered to in order to manufacture the product as per customer requirements. This process also tells where each job will be routed to after the completion of each operation. Furthermore, if there are any non-conforming features of a customer dataset (blueprint requirements of a part type) then the unique gear must undergo what is known as the "QN Rework Process". A "QN" is known as a *Quality Notification* and is initiated by Site Quality during routine inspections between operations. This rework process is a special set of instructions that attempts to fix the defect found on the part. A mantra of the aerospace industry is rework whenever possible, scrap only when necessary. Unlike other industries such as automotive, Collins will attempt to rework instead of scrap whenever they can.

For the arrival rate of new jobs (parts) into the system, an exponential arrival rate was used with each of the different machines acting as a different node in the network. As stated above, each part goes through a certain sequence of operations and each of those operations can be done on certain nodes. Take the following example of baking a batch of cookies:

Imagine someone has four different kinds of cookies: Chocolate chip, double chocolate, white macadamia and mint chocolate. You create a batch of thirty cookies for each type. The process to make a cookie might be: "Make the dough, shape it with cookie cutters, preheat oven, bake cookies". These four steps are each done by a different tool or person. After each step the cookies are inspected by only checking the FIRST cookie in each batch to make sure they are being done according to the recipe. If they are, they move onto the next step. If not, say they don't have enough sugar, they must correct the error to rework it back to the recipe. When we say that

1. Harchol-Balter, M. (2014). *Performance modeling and design of computer systems: queueing theory in action*. New York: Cambridge University Press.

an operation (step) has alternate nodes that simply means that in the case of these cookies that there is more than one cookie cutter or five people are making the dough, or there are 5 ovens that can bake these cookies at a time.

There are five major departments that control the production of landing gear at Collins Aerospace. They are: Site Quality – oversee inspections after ***each*** operation to ensure no measurements are non-conforming, Manufacture Engineering (MFG Eng.) – creates the manufacturing plan (instructions to tell the machine operators how to create the part) and offers dispositions on easily fixable defects in the machining process (known as "Grief" QNs), Material Review Board Engineering (MRB Eng.) – oversees more complex defects that are not easily fixable and gives a disposition on whether or not that defect will affect the structural integrity of the rest of the aircraft and how it can be fixed (known as "MRB" or "Withhold" QNs), Material Review Board Quality Assurance (MRB QA) – these are the liaisons between MFG ENG, MRB ENG, Manufacturing, and Site Quality; they ensure that all QNs raised are done so correctly and route the QN rework to the correct machine, Manufacturing – Is responsible for machining (actually creating) the parts as per the manufacturing process sheets created by MFG Eng.

Finally, if there is a non-conformance the part will go into a subprocess known as the rework process (QN process). An inspector first notifies (raises) the nonconformance in the system and sends the ticket to an MRB QA to have it approved. From there, if it is easily fixable it will be sent to MFG Eng. for disposition, otherwise it will go to MRB Eng. for dispositioning. After the part has been dispositioned it will be fixed by the machine shop and then re-inspected by Quality afterwards to ensure that the problem has been fixed. If Quality says, in the last step,

1. Harchol-Balter, M. (2014). *Performance modeling and design of computer systems: queueing theory in action*. New York: Cambridge University Press.

the defect has been repaired then the QN is closed and the part proceeds with the next operation, otherwise it goes back to Engineering for disposition, reworked by the machine shop (manufacturing) and re-inspected. This loop will continue until the problem is fixed.

III – Assumptions used

In order to keep analysis simple, an M/M/1 queue was used to model the system. Most operations had a list of alternate nodes that could be used to complete an operation. Say if an operation can be completed on nodes *a, b, c, d, e* then that operation would have a processing rate of 5 μ. For the probability routing between nodes, a uniform distribution was used to show that there is an equal among all nodes to complete a given operation. Also note that in the simulation, if a node was free then it immediately took a job that was waiting. A job is marked complete in the system when it successfully completes its last operation. For any dates of events that begin and end on the same day, it was assumed that it took an entire shift (8 hours) to complete that event.

Even though an exponential arrival rate was used, a hard cap was put on the number of parts being manufactured. Evidence from work orders show that in the past two years, 360 parts have been successfully completed and shipped out which equates to roughly 90 jobs in the system for each part type A, B, C, D.

Also, since there was no way to account for some operations, such as heat treat for example, they occurred externally of Collins' network and retrieved the average time spent out

1. Harchol-Balter, M. (2014). *Performance modeling and design of computer systems: queueing theory in action*. New York: Cambridge University Press.

for heat treat from official shipment records and saw that the process time for that external process was 2 weeks.

Another important assumption made had to deal with staff known as “CNC programmers”. CNC programmers are the staff that maintain the programs that the machines/nodes run on. These are the instructions that tell the machines/nodes how to execute the operation. If there is a defect then they will change it so that doesn’t happen again on the next part. Sometimes however, the programmers must update programs that have nothing wrong with them due to receiving new tooling, after having maintenance on the machine, or simply cleaning up a program of any unnecessary code. It was decided that a program will change for a node every month due to the records of the past 30 years showing a similar trend. The consequence of a program changing simply means that any proven node/machine for an operation must be reinspected by Quality in order to reprove the machine. It is because of this assumption that one should expect to see a higher utilization for inspection since this creates a lot of extra work.

Finally, a rate was needed for the probability of a defect occurring after each operation. In 360 parts manufactured over 2 years, there were approximately 1500 QNs raised for those parts. That means on average each part had 4-5 QNs throughout its lifetime being manufactured. If a machine/node was proven by quality meaning that no defects were found during the inspection process after an operation, then it was said there was a 15% chance of every subsequent part going to QN due to human error or handling damage. If a machine/node had not been proven, then there was an 85% chance of a part having a defect and needing to go into the QN rework process. The reason why such a high percentage for defects was used was the fact that so many

1. Harchol-Balter, M. (2014). *Performance modeling and design of computer systems: queueing theory in action*. New York: Cambridge University Press.

QNs existed for only 360 parts. If one were to pick at random any part being made at Collins, you would definitely see *at least* 1 QN attached to it.

IV - Methods used for Analysis and Simulation

Analysis was performed using a Classed Jackson-Network model. This entailed creating a series of classes that mimicked the role of the system of a Jackson-Network to service the large data set. Manual conversion and mathematical operations were deemed infeasible due to the size of the data set.

Given the routing probabilities per class for all nodes in the system, the service times and the external arrival rates, the total arrival rates were determined. Since the classed arrival rate for any node in the system is equal to the sum of its external arrival rate to that node and the product of the routing probabilities to that node with the source nodes respective classed arrival rate, a linear system of equations could be created.

Solving the system was a simple case of inverting the coefficient matrix however the process was made more efficient by the use of the linalg.solve function in the numpy library. This function implements the Lapack _gesv routine for solving a linear system from the based in the inversion method originally considered.

Once solved, the results from the linear system were summed with the external arrival rates to determine the total arrival rate to the node. Since the analysis was performed with a Jackson-Network, each of the nodes was considered an M/M/1 queue and solving for the response time and utilisation for each process was simple. To verify the arrival rates, the analysis was compared to the nominal example found in *Performance Modeling and Design of Computer Systems: Queueing Theory In Action* by Mor Harchol-Balter on page 323.

1. Harchol-Balter, M. (2014). *Performance modeling and design of computer systems: queueing theory in action*. New York: Cambridge University Press.

For class 1: $r_1(1) = 3; \quad P_{12}^{(1)} = 1; \quad P_{23}^{(1)} = 1; \quad P_{3,\text{out}}^{(1)} = 1$

For class 2: $r_1(2) = 4; \quad P_{13}^{(2)} = 1; \quad P_{34}^{(2)} = 1; \quad P_{4,\text{out}}^{(2)} = 1$

For class 3: $r_2(3) = 5; \quad P_{23}^{(3)} = 1; \quad P_{34}^{(3)} = 1; \quad P_{4,\text{out}}^{(3)} = 1$

For class 4: $r_3(4) = 6; \quad P_{3,\text{out}}^{(4)} = 1$

Figure 1: Nominal values taken from the textbook

$\lambda_3(1) = \lambda_2(1) = \lambda_1(1) = r_1(1) = 3$ jobs/sec

$\lambda_4(2) = \lambda_3(2) = \lambda_1(2) = r_1(2) = 4$ jobs/sec

$\lambda_4(3) = \lambda_3(3) = \lambda_2(3) = r_2(3) = 5$ jobs/sec

$\lambda_3(4) = r_3(4) = 6$ jobs/sec

Figure 2: Actual arrival rates in our network

Much of the analysis was automated to receive formatted data to minimize data input errors and was tested using test data from *Performance Modeling and Design of Computer Systems: Queueing Theory In Action* by Mor Harchol-Balter.

The network was simulated using a series of first-in-first-out (FIFO) nodes that represent each of the processes within the system. Each of these nodes independently could arrive jobs to the node and depart them after the allotted service time of the job had passed. The simulation created a new FIFO node for each of the machines used during the processing of a part and arrival streams for each of our 4 part types for which the arrivals were exponentially distributed at a rate of 0.00365.

Each of the jobs holds information in regards to its class. Jobs have a specific set of processes that need to be done on it in its lifetime before departing the system. These jobs can be done on one or more different nodes in the network depending on the class along with having varying service times depending on the class. The data is pulled from our processing data and given to each class of jobs.

The simulation functions by having a running clock that is set to run for a given amount of time and increments by checking for the closest event in time which can be either an arrival of a job to a node or a departure of a job from a node. Each of the jobs holds information for its routing in regards to its class as different types of jobs require different processes to be done on it. At any given routing, there can

1. Harchol-Balter, M. (2014). *Performance modeling and design of computer systems: queueing theory in action*. New York: Cambridge University Press.

be multiple machines that a job is able to be routed to. For the simulation, jobs are conditionally routed to the machine with the shortest length queue time. In addition to the conditionally routing of the required process, jobs can also be routed to a set of QN processes given that the FIFO node that the job arrived at requires it. To match the process of Collin's, every month there is a higher possibility for that machine to need to send a job to QN while it is lower at all other times. In the simulation those probabilities are set to 0.05 and 0.01 respectively.

With the available data for the QN process, the routing between different processes within QN is extremely inconsistent and can vary widely from ordering to service times to the names of the node itself. Another challenge comes in that it is unclear which QN process in the data should be the final process for a part and therefore return it back to its normal processing. Our solution was to create a FIFO node for every process found in the data regardless of any formatting error or equivalencies in nodes as it is difficult to determine such a large data set. Additionally, the final nodes that jobs were processed in the QN data were treated as departure nodes and given their respective departure probabilities regardless of whether a different node should have in practice been the departing node.

Upon returning from QN the job is then routed through each of it's required nodes and any further QN processing until either the simulation timer ends or the required set of processes done of the job has been exhausted.

With the simulation running, the FIFO nodes themselves are set to track the response times of nodes waiting to be serviced, the utilization of the the node, and the average service time that the node has experienced through the lifetime of the sim.

1. Harchol-Balter, M. (2014). *Performance modeling and design of computer systems: queueing theory in action*. New York: Cambridge University Press.

V - Findings

Earlier, listed were the 5 major parties involved in the production of a landing gear with the parties being: MRB QA, MRB Eng., MFG Eng., Site Quality and Manufacturing (Machine Shop). Here are the results for each of the 5 major parties in terms of utilisation, service time, and response time for QN processes:

Utilisation

MRB QA: 84.17%

Site Quality: 70.81%

MRB Engineering: 25.40%

MFG Engineering: 5.31%

Machine Shop: 92.93%

Average Service Times and Rates (in hours)

MRB QA: 22.02746 | 0.019311388930944768 jobs/hr

Site Quality: 66.4014 - 155.2941 | 0.01505992343 - 0.00643939467 jobs/hr

MRB Engineering: 25.5800 | 0.016376750059752566 jobs/hr

MFG Engineering: 51.6944 | 0.00975316661840254 jobs/hr

Machine Shop: 215.5626 | 0.645587213342599 jobs/hr

Average Response Time (in hours)

MRB QA: 63.1064

Site Quality: 77.5579 - 162.2325

MRB Engineering: 27.4705

MFG Engineering: 53.5801

*Machine Shop: 1626.6792

1. Harchol-Balter, M. (2014). *Performance modeling and design of computer systems: queueing theory in action*. New York: Cambridge University Press.

Looking at the numbers from the simulation, this is expected behaviour. Every part made in the plant must go through Site Quality for routine inspection at every step of the machining process. Since the time is in hours, looking at the data contained within the records of Collins' databases the expected average wait time for dispositions from MRB and MFG engineers are at 1 day and 2 days respectively. The only number that is not behaving normally is the Machine Shop (Also known as Manufacturing). With the current average response time, that would be the equivalent of 67.75 days! The reason why this is so high is because that part of the QN process is unstable. The analysis shows that this part of the network is unstable. What this means is that the number of jobs waiting to be done by the machine shop at any of the machines will continue to rise. The reason why Site Quality has ranges for their service and response times is that when pulling information from SAP (The database Collins uses for QNs) There are three different headings that can be assigned to an Inspector. For the utilisation rate, all of the numbers were added together for the total utilisation of (Site) Quality. For the service and response times a range from the highest and lowest numbers was used. These numbers imply that a job waits between 5 to 7 days before an inspector looks at it. This is precisely how it behaves in the real world. It is important to note though that even though Quality has a high utilization and response time, they are simultaneously one of the causes *and not one of the causes* of the bottleneck. They are considered a bottleneck in the current system but if one were to look at the system they would see it is flawed. An 85% rate of going into the rework process implies that almost all, *if not all* parts are not being made properly and require additional reworks and inspections. Quality, much like a programmer's test cases are used to only find problems, they themselves are not the cause of the error. Further research would be required to identify why it is that there is such an abnormally high number of non-conformities that is leading to the high number of QNs being raised.

1. Harchol-Balter, M. (2014). *Performance modeling and design of computer systems: queueing theory in action*. New York: Cambridge University Press.