Computing Infrastructures
Exercises

Christian Rossi

Academic Year 2023-2024

Abstract

The course topics are:

- Hardware infrastructure of datacenters:
  - Basic components, rack structure, cooling.
  - Hard Disk Drive and Solid State Disks.
  - RAID architectures.
  - Hardware accelerators.
- Software infrastructure of datacenters:
  - Virtualization: basic concepts, technologies, hypervisors and containers.
  - Computing Architecture: Cloud, Edge and Fog Computing.
  - Infrastructure, platform and software-as-a-service.
- Methods:
  - Scalability and performance of datacenters: definitions, fundamental laws, queuing network theory basics.
  - Reliability and availability of datacenters: definitions, fundamental laws, reliability block diagrams.

Contents

CHAPTER 1

Dependability

1.1 Exercise one

A pacemaker for the heart has a failure rate of $\lambda = 0.25 \cdot 10^{-8}$ per hour.

1. Compute the mean time to failure.
2. Compute the probability that it fails during the first five years of operation.

Solution

1. The average time to failure is calculated as:

   $$\text{MTTF} = \frac{1}{\lambda} = \frac{1}{0.25 \cdot 10^{-8}} = 4 \cdot 10^{8} \text{ hours} \approx 45.662 \text{ years}$$

   Hence, the average time to failure of the heart pacemaker is approximately 45.6 years.

2. The reliability of the system can be expressed as:

   $$R(t) = e^{-\lambda t}$$

   Given λ as specified and a duration of five years, we have:

   $$R(5) = e^{-5\lambda} = e^{-1.25 \cdot 10^{-8}} = 0.9999999875$$

   Subsequently, the probability of failure during the initial five years of operation is calculated as:

   $$F(5) = 1 - R(5) = 1 - 0.9999999875 = 1.25 \cdot 10^{-8}$$

1.2 Exercise two

Let's consider a generic component D. We are tasked with determining the minimum integer value for the mean time to failure of D to ensure that $R_{\text{D}}(t) \geq 0.96$ at five days.

Solution

The reliability is defined as:

$$R(t) = e^{-\lambda t} = e^{-\frac{t}{\mathrm{MTTF}}}$$

We need to satisfy:

$$R_{\mathrm{D}}(5) \geq 0.96$$

Which is equivalent to:

$$e^{-\frac{5\mathrm{days}}{\mathrm{MTTF}}} \geq 0.96$$

Thus, we have:

$$\mathrm{MTTF} \geq -\frac{5\mathrm{days}}{\ln(0.96)} \to \mathrm{MTTF} \geq 122.5 \text{ days}$$

Therefore, the minimum mean time to failure required to meet the given conditions is 122.5 days.

1.3 Exercise three

A smartphone manufacturer determines that their products have a mean time to failure of 59 years in normal use. They want to estimate how long a warranty should be set if no more than 5% of the items are to be returned for repair.

Solution

We seek a time t_w such that:

$$1 - R(t_w) = 0.05 \to R(t_w) = 0.95$$

Given the mean time to failure, we can express the reliability as:

$$R(t) = e^{-\lambda t} \to e^{-\frac{t}{\mathrm{MTTF}}}$$

Substituting, we get:

$$R(t_w) = e^{-\frac{t_w}{\mathrm{MTTF}}} = 0.95 \to t_w = -59 \ln(0.95) = 3.026 \text{ years}$$

Hence, the warranty should be set for approximately 3 years to ensure that no more than 5% of the items are returned for repair.

1.4 Exercise four

A complex system has a failure rate of $\lambda = 0.25 \cdot 10^{-4}$ per hour and a mean time to repair of 72 hours in normal use.

1. Compute the steady-state availability.
2. If mean time to repair is increased to 120 hours, compute the failure rate that can be tolerated without decreasing the availability of the system.

Solution

1. To calculate the steady-state availability, we first find the mean time to failure:

$$\text{MTTF} = \frac{1}{\lambda} = \frac{1}{0.25 \cdot 10^{-4}} = 40000 \text{ hours}$$

Then, the availability is computed as:

$$A = \frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}} = \frac{40000}{40000 + 72} = 0.9982$$

2. If we increase the mean time to repair while maintaining the same availability of 0.9982, we first find the new minimum mean time to failure:

$$A = \frac{\text{MTTF}_{new}}{\text{MTTF}_{new} + \text{MTTR}} \rightarrow \text{MTTF}_{new} = -\frac{A \cdot \text{MTTR}}{A - 1} \rightarrow \text{MTTF}_{new} = 66666.66 \text{ hours}$$

From this, we derive the new failure rate:

$$\lambda_{new} = \frac{1}{\text{MTTF}_{new}} = \frac{1}{66666.66} = 1.5 \cdot 10^{-5}$$

1.5 Exercise five

Consider a server architecture composed of three main components: CPU, memory, and hard drive. Each component has a constant failure rate of $\frac{1}{64}$, $\frac{1}{58}$ and $\frac{1}{28}$ per year, respectively, and failures are assumed to be independent events.

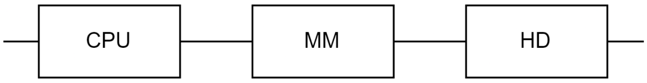
1. Visualize the reliability block diagram of the server architecture.

2. Calculate the mean time to failure for the server.

3. Determine the reliability of the server over a three-year mission.

Solution

1. The reliability block diagram of the server architecture is depicted below:



Given that the components are in series, the total reliability is the sum of the reliabilities of each component:

$$R(t) = R_{CPU}(t) + R_{memory}(t) + R_{hard\,disk}(t) = e^{-\lambda_{CPU}t} + e^{-\lambda_{memory}t} + e^{-\lambda_{hard\,disk}t}$$

2. As the components are in series, the total failure rate is the sum of individual failure rates:

$$\lambda_{tot} = \lambda_{CPU} + \lambda_{memory} + \lambda_{hard\,disk} = \frac{1}{64} + \frac{1}{58} + \frac{1}{28} = \frac{891}{12992}$$

The mean time to failure is the inverse of the total failure rate:

$$\text{MTTF}_{tot} = \frac{1}{\lambda_{tot}} = \frac{1}{\frac{891}{12992}} = 14.58 \text{ years}$$

3. The reliability for a three-year mission is:

$$R_{tot}(3 \text{ years}) = e^{-3\lambda_{tot}} = 0.814$$

1.6 Exercise six

A computer system is engineered with a failure rate of one fault every five years under typical usage conditions. This system lacks fault tolerance capabilities, meaning it ceases functioning upon encountering its initial fault.

1. Compute the mean time to failure of the system.
2. Compute the probability that the system will fail during its first year of operation.
3. The standard warranty for the system covers 2 years of operation. However, the vendor aims to provide extended insurance against failures for the initial 5 years of operation, for which they plan to charge an additional fee. The vendor intends to charge 20$ for every 1% decrease in reliability to offer this insurance. Compute how much should the vendor charge for such an insurance.

Solution

1. The mean time to failure for the system is calculated as the inverse of the failure rate:

$$\text{MTTF} = \frac{1}{\lambda} = \frac{1}{\frac{1}{5}} = 5 \text{ years}$$

2. The failure probability within a given time frame is given by:

$$F(t) = 1 - R(t)$$

For the first year:

$$F(1) = 1 - R(1) = 1 - e^{-\lambda} = 1 - 0.818 = 0.18$$

3. We compute the reliability at 2 and 5 years:

$$R(2 \text{ year}) = e^{-2\lambda} = 0.67$$

$$R(5 \text{ year}) = e^{-2\lambda} = 0.37$$

This yields a reliability drop of 0.3. Since the vendor wants to charge 20$ for each 1% drop in reliability, the cost is:

$$\text{cost} = \frac{0.3}{0.01} = 30 \cdot 20\$ = 600\$$$

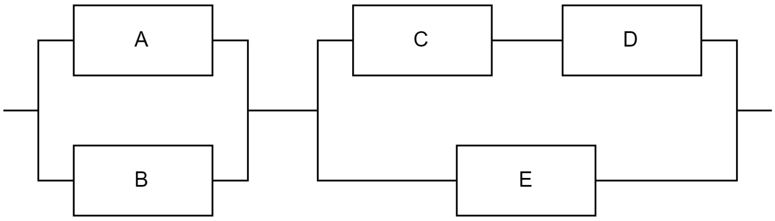
1.7 Exercise seven

A system comprising five modules (A, B, C, D, and E) is designed to function properly under the following conditions: either modules A or B operate correctly, and concurrently modules C and D operate correctly, or alternatively, module E operates correctly.

1. Draw the reliability block diagram of the system.
2. Find the expression for the reliability of the system.
3. Given that the mean time to failure for modules A and B is 3412 hours, and for modules C, D, and E is 1245 hours, we aim to calculate the reliability of the system after 1 month.

Solution

1. The reliability block diagram of the system is depicted below:



2. The blocks A and B are in parallel, hence the total reliability is computed as:

$$R_{\mathrm{AB}} = 1 - (1 - R_{\mathrm{A}})\,(1 - R_{\mathrm{B}})$$

The blocks C and D are in series, therefore the total reliability is:

$$R_{\mathrm{CD}} = R_{\mathrm{C}} R_{\mathrm{D}}$$

Now, with the new block CD in parallel with E, the reliability becomes:

$$R_{\mathrm{CDE}} = 1 - (1 - R_{\mathrm{CD}})\,(1 - R_{\mathrm{E}})$$

Finally, the entire system's reliability is the product of the reliabilities of blocks AB and CDE:

$$R_s = [1 - (1 - R_{\mathrm{A}})\,(1 - R_{\mathrm{B}})]\,[1 - (1 - R_{\mathrm{CD}})\,(1 - R_{\mathrm{E}})]$$

3. We can compute the monthly mean time to failures:

$$\mathrm{MTTF}_{\mathrm{A}} = \mathrm{MTTF}_{\mathrm{B}} = 4,74 \text{ months}$$

$$\mathrm{MTTF}_{\mathrm{C}} = \mathrm{MTTF}_{\mathrm{D}} = \mathrm{MTTF}_{\mathrm{E}} = 1,73 \text{ months}$$

The monthly failure rates are the inverses of the mean time to failures:

$$\lambda_{\mathrm{A}} = \lambda_{\mathrm{B}} = \frac{1}{4,74 \text{ months}} = 0.21$$

$$\lambda_{\mathrm{C}} = \lambda_{\mathrm{D}} = \lambda_{\mathrm{E}} = \frac{1}{1,73 \text{ months}} = 0.578$$

We can now compute the reliability of each block:

$$R_{\mathrm{A}}(1) = R_{\mathrm{B}}(1) = e^{-\lambda t} = e^{-1 \cdot 0.21} = 0.81$$

$$R_{\mathrm{B}}(1) = R_{\mathrm{C}}(1) = R_{\mathrm{D}}(1) = e^{-\lambda t} = e^{-1 \cdot 0.578} = 0.56$$

From the previously derived formula, we find:

$$R_s(1) = [1 - (1 - R_{\mathrm{A}}(1))\,(1 - R_{\mathrm{B}}(1))]\,[1 - (1 - R_{\mathrm{CD}}(1))\,(1 - R_{\mathrm{E}}(1))] = 0.67$$

Specifically, we have:

- $R_{\mathrm{AB}} = 1 - (1 - R_{\mathrm{A}})\,(1 - R_{\mathrm{B}}) = 0.96.$
- $R_{\mathrm{CD}} = R_{\mathrm{C}} R_{\mathrm{D}} = 0.31.$
- $R_{\mathrm{CDE}} = 1 - (1 - R_{\mathrm{CD}})\,(1 - R_{\mathrm{E}}) = 0.70.$

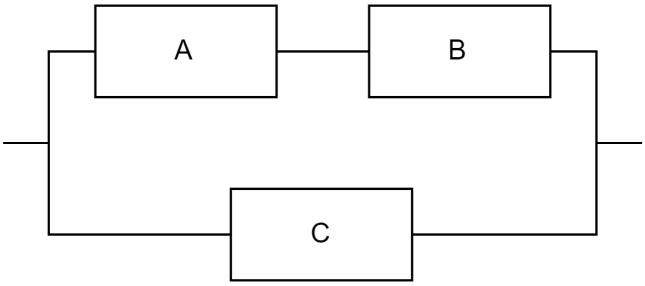
1.8 Exercise eight

The system comprises three modules (A, B, and C) and is designed to function correctly if both A and B operate correctly or if C operates correctly. The availabilities of the modules are as follows: $A_{\mathrm{A}} = 0.97$, $A_{\mathrm{B}} = 0.92$, and $A_{\mathrm{D}} = 0.95$.

1. Illustrate the reliability block diagram of the system.
2. Compute the system's availability.

Solution

1. The reliability block diagram of the system is depicted below:



2. To calculate the availability of the system, we first determine the availability of blocks A and B in series:

$$A_{\mathrm{AB}} = A_{\mathrm{A}} A_{\mathrm{B}} = 0.97 \cdot 0.92 = 0.89$$

Subsequently, with block AB in parallel with C, the overall availability becomes:

$$A_s = 1 - (1 - A_{\mathrm{AB}})(1 - A_{\mathrm{C}}) = 1 - (0.11)(0.04) = 0.9956$$

1.9 Exercise nine

A system comprises four non-redundant components, with a 1-year reliability of 0.92. A new version of the system incorporates a novel feature and utilizes six non-redundant components. Compute the 1-year reliability of the new system, assuming all components have the same mean time to failure.

Solution

The reliability of the initial system is:

$$R_{old}(1 \text{ year}) = \left(e^{-\lambda}\right)^4 = e^{-4\lambda}$$

Given this expression equals 0.92, we can determine the value of the failure rate:

$$e^{-4\lambda} = 0.92 \rightarrow \lambda = \frac{\ln(0.92)}{-4} = 0.02$$

From the failure rate, we can derive the mean time to failure:

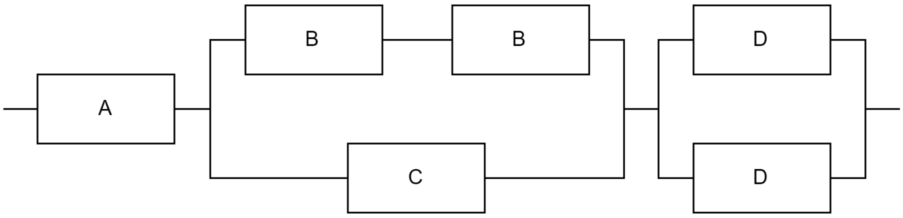
$$\mathrm{MTTF} = \frac{1}{\lambda} = \frac{1}{0.02} = 47.97 \text{ years}$$

In the new scenario, the reliability becomes:

$$R_{old}(1 \text{ year}) = \left(e^{-\lambda}\right)^6 = e^{-6\lambda} = 0.88$$

1.10 Exercise ten

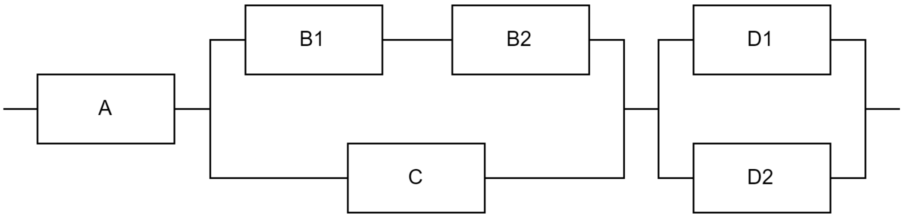
A system is composed of components arranged in the following reliability block diagram:



1. Identify all possible configurations of components that may fail without causing the entire system to fail.
2. Calculate MTTF_B knowing that $R(t) = 83\%$ at two years.
3. Compute the availability of the entire system given that $\text{MTTF}_\text{A} = \text{MTTF}_\text{D} = 1$ year, $\text{MTTF}_\text{C} = 14$ months, and $\text{MTTR} = 21$ days for all components.

Solution

1. Renaming the blocks in the reliability block diagram as:



The pairs of blocks that may fail without causing the entire system to fail are:

- B1 and D1.
- B1 and D2.
- B2 and D1.
- B2 and D2.
- C and D1.
- C and D2.

2. The reliability of each block B is calculated as:

$$R_B(2 \text{ years}) = e^{-\frac{2}{\text{MTTF}_\text{B}}}$$

Given this expression equals 0.83, we solve for MTTF:

$$e^{-\frac{2}{\text{MTTF}_\text{B}}} = 0.83 \rightarrow \text{MTTF} = -\frac{2}{\ln(0.83)} = 10.73 \text{ years}$$

3. The mean time to failure of block C in years is:

$$\mathrm{MTTF_C} = \frac{14 \text{ months}}{12 \text{ months}} = 1.17 \text{ years}$$

Similarly, the mean time to recovery is:

$$\mathrm{MTTR} = \frac{21 \text{ days}}{365 \text{ days}} = 0.058 \text{ years}$$

We can compute the availability for each block:

- $A_\mathrm{A} = A_\mathrm{D} = \frac{\mathrm{MTTF}}{\mathrm{MTTF+MTTR}} = \frac{1}{1+0.058} = 0.95$
- $A_\mathrm{C} = \frac{\mathrm{MTTF}}{\mathrm{MTTF+MTTR}} = \frac{1.17}{1.17+0.058} = 0.95$
- $A_\mathrm{B} = \frac{\mathrm{MTTF}}{\mathrm{MTTF+MTTR}} = \frac{10.73}{10.73+0.058} = 0.99$

Now, we combine the two blocks B into a single block:

$$A_\mathrm{BB} = A_\mathrm{B} A_\mathrm{B} = 0.99 \cdot 0.99 = 0.98$$

Then, we have the parallel connection between the new block and C:

$$A_\mathrm{BBC} = 1 - (1 - A_\mathrm{BB})\,(1 - A_\mathrm{C}) = 1 - (0.02)(0.05) = 0.999$$

Similarly, for the parallel connection between the two blocks D:

$$A_\mathrm{DD} = 1 - (1 - A_\mathrm{D})\,(1 - A_\mathrm{D}) = 1 - (0.05)(0.05) = 0.997$$

Finally, the total availability of the system can be computed as the series of A, BBC, and DD:

$$A_s = 1 - (1 - A_\mathrm{A})\,(1 - A_\mathrm{BBC})\,(1 - A_\mathrm{DD}) = 1 - (0.05)(0.001)(0.003) = 0.946$$

1.11 Exercise twelve

In the C-5 aircraft, there are 12 identical AC generators, and at least 9 of them must be operational for the aircraft to complete its mission. Failures follow an exponential distribution with a failure rate of 0.01 failure per hour. Compute the reliability of the generator system over a ten-hour mission in case the switch is perfect.

Solution

For a system composed of n identical replicas where at least r replicas must function for the entire system to operate correctly, the system reliability is given by:

$$R_s(t) = R_V \sum_{i=r}^{n} R_c^i (1 - R_c)^{n-1} \binom{n}{i}$$

Here:

- R_s is the system reliability.

- R_c is the component reliability.
- R_v is the voter reliability.
- n is the number of components.
- r is the minimum number of components that must function.

For each generator over a ten-hour mission, the reliability is:

$$R_m(10) = e^{-\lambda t} = e^{-0.01 \cdot 10} = 0.9048374$$

Substituting into the formula, we get:

$$R_s = \sum_{i=9}^{1} 2R_m^i \left(1 - R_m\right)^{12-i} \binom{12}{i} = 165R_m^{12} + 540R_m^{11} + 549R_m^1 0 + 220R_m^9 = 0.9782773$$

CHAPTER 2

Redundant Array of Independent Disks

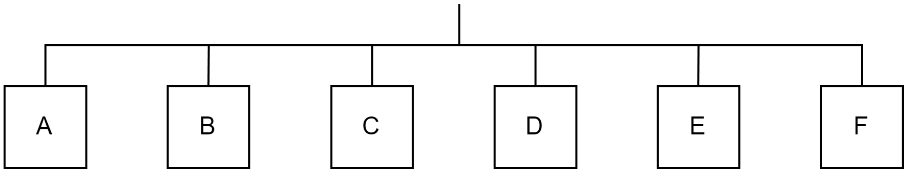
2.1 Exercise one

Let's explore the total storage capacity of a system consisting of six disks, each with a capacity of 1 TB, under various RAID configurations:

1. RAID 0
2. RAID 1
3. RAID 01
4. RAID 10
5. RAID 5
6. RAID 6

Solution

1. In RAID 0, all disks are utilized as a single unit:

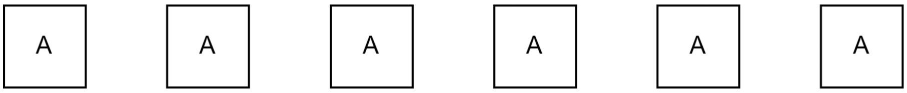


Hence, the total capacity is the sum of individual disk capacities:

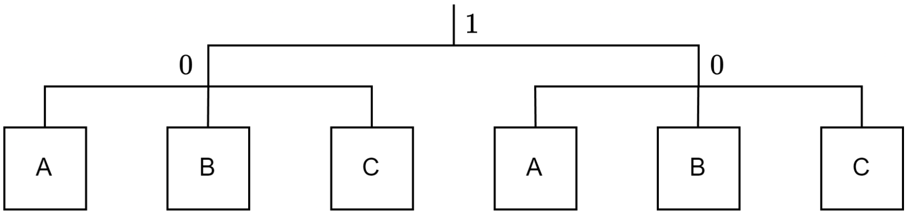
$$S_C = 6 \cdot 1 \text{ TB} = 6 \text{ TB}$$

2. RAID 1 duplicates data across disks for high reliability but lower storage:

$$S_C = 1 \text{ TB}$$



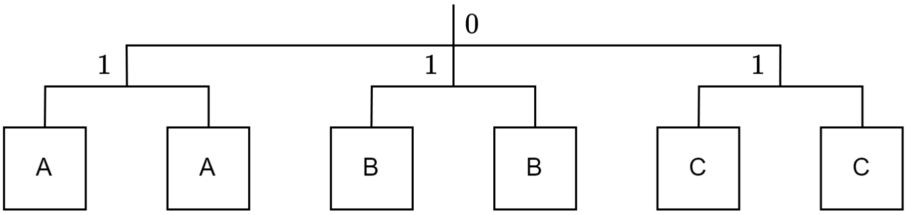
3. RAID 01 involves striping followed by mirroring, resulting in two sets of mirrored blocks:



Thus, the total capacity becomes:

$$S_C = \frac{6}{2} \cdot 1 \text{ TB} = 3 \text{ TB}$$

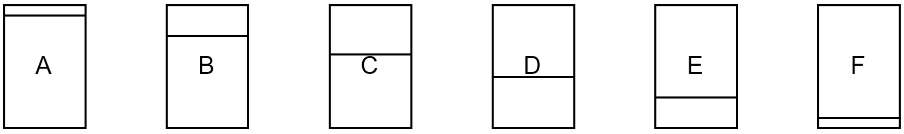
4. RAID 10 employs mirroring followed by striping, creating three pairs of mirrored disks:



Therefore, the total capacity is:

$$S_C = \frac{6}{2} \cdot 1 \text{ TB} = 3 \text{ TB}$$

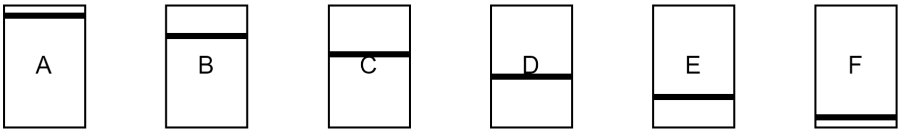
5. RAID 5 utilizes parity blocks across disks, requiring removal of one disk:



This yields:

$$S_C = (N-1) \cdot 1 \text{ TB} = 5 \text{ TB}$$

6. RAID 6 employs two parity blocks per disk, necessitating removal of two disks:



Resulting in:

$$S_C = (N-2) \cdot 1 \text{ TB} = 4 \text{ TB}$$

2.2 Exercise two

Let's analyze the mean time to failure (MTTF) of a RAID 0 setup with the following specifications:

- $N = 5$ disks.
- Mean Time to Repair (MTTR) of eight hours.
- Mean Time to Failure (MTTF) of one disk is 1600 days.

Solution

In RAID 0, all elements are in series, so the MTTF of each disk divided by the number of disks equals the MTTF of the RAID 0 setup:

$$\text{MTTF}_{\text{RAID 0}} = \frac{\text{MTTF}}{\text{N}} = \frac{1600}{5} = 320 \text{ days}$$

2.3 Exercise three

Let's calculate the mean time to failure (MTTF) of a RAID 1 setup with the following specifications:

- $N = 2$ disks.
- Mean Time to Repair (MTTR) of eight days.
- Mean Time to Failure (MTTF) of one disk is 1800 days.+

Solution

The failure rate of the disks is given by:

$$\lambda = \frac{N}{\text{MTTF}} = \frac{2}{1800} = \frac{1}{900}$$

The failure rate of the RAID 1 setup is then:

$$\lambda_{\text{RAID 1}} = \lambda \cdot \left(\frac{1}{\text{MTTF}} \cdot \text{MTTR}\right) = \frac{2 \cdot \text{MTTR}}{\text{MTTF}^2}$$

Therefore, the mean time to failure is:

$$\text{MTTF}_{\text{RAID 1}} = \frac{1}{\lambda_{\text{RAID 1}}} = \frac{1800^2}{2 \cdot 8}$$

2.4 Exercise four

Let's calculate the mean time to failure (MTTF) of a RAID 10 and RAID 01 setup with the following specifications:

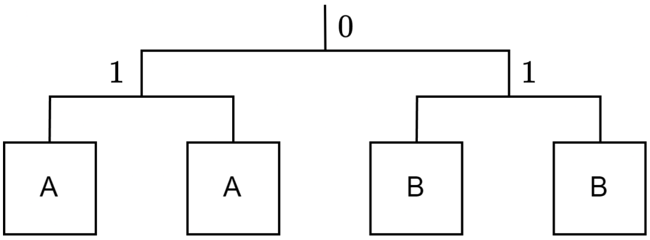
- $N = 4$ disks.
- Mean Time to Repair (MTTR) of three days.
- Mean Time to Failure (MTTF) of a single disk is 1400 days.

Compte the Mean Time to Failure (MTTF) in case of:

1. RAID 10.
2. RAID 01.

Solution

1. For the RAID 10 setup:



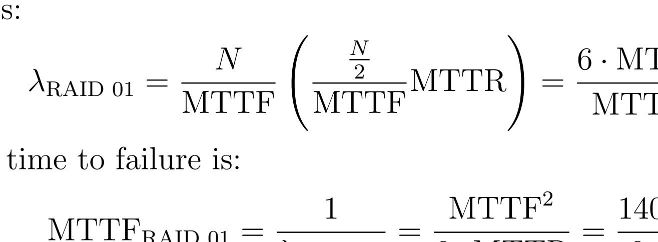
The failure rate is:

$$\lambda_{\text{RAID 10}} = \frac{N}{\text{MTTF}}\left(\frac{1}{\text{MTTF}}\text{MTTR}\right) = \frac{4\cdot\text{MTTR}}{\text{MTTF}^2}$$

Thus, the mean time to failure is:

$$\text{MTTF}_{\text{RAID 10}} = \frac{1}{\lambda_{\text{RAID 10}}} = \frac{\text{MTTF}^2}{4\cdot\text{MTTR}} = \frac{1400^2}{3\cdot 4}$$

2. For the RAID 01 setup:



The failure rate is:

$$\lambda_{\text{RAID 01}} = \frac{N}{\text{MTTF}}\left(\frac{\frac{N}{2}}{\text{MTTF}}\text{MTTR}\right) = \frac{6\cdot\text{MTTR}}{\text{MTTF}^2}$$

Hence, the mean time to failure is:

$$\text{MTTF}_{\text{RAID 01}} = \frac{1}{\lambda_{\text{RAID 10}}} = \frac{\text{MTTF}^2}{6\cdot\text{MTTR}} = \frac{1400^2}{6\cdot 4}$$

2.5 Exercise five

Let’s calculate the mean time to failure (MTTF) of a RAID 10 and RAID 01 setup with the following specifications:

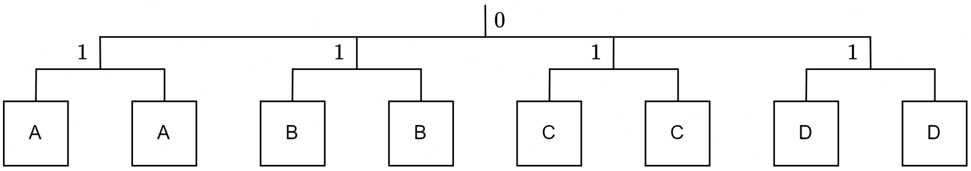
- $N = 8$ disks.
- Mean Time to Repair (MTTR) of four days.
- Mean Time to Failure (MTTF) of a single disk is 2200 days.

Compte the Mean Time to Failure (MTTF) in case of:

1. RAID 10.
2. RAID 01.

Solution

1. For the RAID 10 setup:



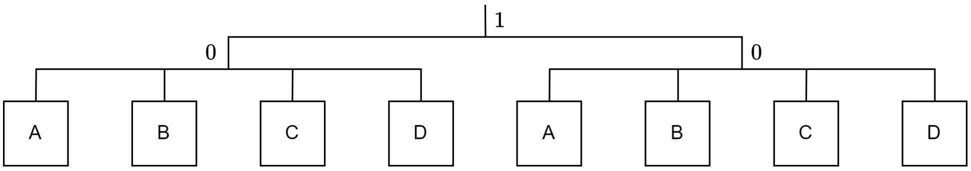
The failure rate is:

$$\lambda_{\text{RAID 01}} = \frac{N}{\text{MTTF}} \left(\frac{1}{\text{MTTF}} \text{MTTR} \right) = \frac{8 \cdot \text{MTTR}}{\text{MTTF}^2}$$

Hence, the mean time to failure is:

$$\text{MTTF}_{\text{RAID 10}} = \frac{1}{\lambda_{\text{RAID 10}}} = \frac{\text{MTTF}^2}{8 \cdot \text{MTTR}} = \frac{2200^2}{8 \cdot 4}$$

2. For the RAID 01 setup:



The failure rate is:

$$\lambda_{\text{RAID 01}} = \frac{N}{\text{MTTF}} \left(\frac{\frac{N}{2}}{\text{MTTF}} \text{MTTR} \right) = \frac{32 \cdot \text{MTTR}}{\text{MTTF}^2}$$

Thus, the mean time to failure is:

$$\text{MTTF}_{\text{RAID 01}} = \frac{1}{\lambda_{\text{RAID 10}}} = \frac{\text{MTTF}^2}{32 \cdot \text{MTTR}} = \frac{2200^2}{32 \cdot 4}$$

2.6 Exercise six

A system administrator needs to utilize a stock of disks with the following characteristics:

- Mean Time To Failure (MTTF) of 800 days.
- Mean Time To Repair (MTTR) of twenty days.

The desired operational lifetime of the system is 3 years. Calculate the maximum number of disks that can be used in RAID 01 to ensure a mean time to data loss greater than the system's lifetime.

Solution

We know that:

$$\text{MTTF}_{\text{RAID 01}} = \frac{2 \cdot \text{MTTF}^2}{N^2 \cdot \text{MTTR}}$$

Thus, we have:

$$N = \sqrt{\frac{2 \cdot \text{MTTF}^2}{\text{MTTF}_{\text{RAID 01}} \cdot \text{MTTR}}} = \sqrt{\frac{2 \cdot 800^2}{20 \cdot 3 \cdot 365}} \approx 7.65$$

The correct answer is six, as in RAID 01, an even number of disks is needed, which must be less than $N = 7.65$.

2.7 Exercise seven

Let's consider the following RAID 5 setup:

- $N = 4$ disks.
- Mean Time To Repair (MTTR) of three days.
- Mean Time To Failure (MTTF) of one disk is 1000 days.

Calculate the Mean Time To Failure (MTTF) of this setup.

Solution

The failure rate of the RAID 5 array is given by:

$$\lambda_{\text{RAID 5}} = \frac{N}{\text{MTTF}} \cdot \left(\frac{N-1}{\text{MTTF}} \cdot \text{MTTR} \right) = \frac{12 \cdot \text{MTTR}}{\text{MTTF}^2}$$

Thus, the mean time to failure is:

$$\text{MTTF}_{\text{RAID 5}} = \frac{1}{\lambda_{\text{RAID 5}}} = \frac{1000^2}{12 \cdot 3}$$

2.8 Exercise eight

Let's examine the following RAID 6 setup:

- $N = 5$ disks.
- Mean Time To Repair (MTTR) of two days.
- Mean Time To Failure (MTTF) of one disk is 1100 days.

Compute the Mean Time To Failure (MTTF) of this setup.

Solution

The failure rate of the RAID 6 array is given by:

$$\lambda_{\text{RAID 6}} = \frac{N}{\text{MTTF}} \cdot \left(\frac{N-1}{\text{MTTF}} \cdot \text{MTTR}\right) \cdot \left(\frac{N-2}{\text{MTTF}} \cdot \frac{\text{MTTR}}{2}\right) = \frac{60 \cdot \text{MTTR}^2}{2 \cdot \text{MTTF}^3}$$

Thus, the mean time to failure is:

$$\text{MTTF}_{\text{RAID 6}} = \frac{2 \cdot 1100^3}{60 \cdot 2^2}$$

2.9 Exercise nine

In a RAID 6 system, eight 3 TB drives are used to store data along with the required parity bits. Compute how many identical drives would be required to construct a RAID 10 system with the same capacity. For the RAID 1 configuration, consider a single replica of data blocks.

Solution

The storage capacity for RAID 6 is:

$$S_C^{\text{RAID 6}} = (N-2) \cdot 3 \text{ TB} = 6 \cdot 3 \text{ TB} = 18 \text{ TB}$$

The storage capacity for RAID 10 is:

$$S_C^{\text{RAID 10}} = \frac{N}{2} \cdot 3 \text{ TB} = 18 \text{ TB}$$

Hence, we need twelve drives.

CHAPTER 3

Performance

3.1 Exercise one

A server completed 80 jobs in 10 minutes with a measured utilization of 75%. Determine:

1. The throughput of the system.
2. The average service time.
3. The busy time of the system.
4. The average response time.

Solution

Given:

$$C = 80 \qquad T = 10 \qquad U = 0.75$$

1. Throughput:

$$X = \frac{C}{T} = \frac{80}{10} = 8\,\frac{\text{job}}{\text{min}}$$

2. Average Service Time:

$$B = UT = 0.75 \cdot 10 = 7.5 \text{ min}$$

3. Busy Time:

$$S = \frac{B}{C} = \frac{7.5}{80} = 5.625 \text{ min}$$

4. Average Response Time: unable to compute without additional data.

3.2 Exercise two

A processing system is, on average, handling 35 requests simultaneously and can complete 28 jobs in one hour. Determine:

1. The throughput of the system:

2. The average response time.

3. The utilization of the system.

Solution

Given:

$$N = 35 \qquad T = 1 \qquad C = 28$$

1. Throughput:

$$X = \frac{C}{T} = \frac{28}{1} = 28\,\frac{\text{job}}{\text{hour}}$$

2. Average Response Time:

$$R = \frac{N}{X} = \frac{35}{28} = 1.25\,\text{hours}$$

3. Utilization: unable to compute without additional data.

3.3 Exercise three

The busy time of a virtual machine is 3 hours every 4 hours. Its utilization is 80%, and on average, it is serving 5 jobs simultaneously. Determine:

1. The system throughput.

2. The average response time.

Solution

Given:

$$B = 3 \qquad T = 4 \qquad U = 0.8 \qquad N = 5$$

The utilization calculated from busy time and total time does not match the given utilization:

$$U = \frac{B}{T} = \frac{3}{4} = 0.75 \neq 0.8$$

The given data is inconsistent; therefore, this exercise cannot be solved.

3.4 Exercise four

The busy time of a virtual machine is 3 hours every 4 hours. In that time, it has completed 15 jobs, and on average, it is serving 5 jobs simultaneously. Determine:

- The system throughput.

- The average response time.

Solution

Given:

$$B = 3 \qquad T = 4 \qquad C = 15 \qquad N = 5$$

1. Given:

$$X = \frac{C}{T} = \frac{15}{4} = 3.75\,\frac{\text{jobs}}{\text{hour}}$$

2. Average Response Time:

$$R = \frac{N}{X} = \frac{5}{3.75} = 80 \text{ min}$$

3.5 Exercise five

In a course, there are 100 students. Each student studies for a week, then sends an email to the professor, waits for an answer, studies one more week, and then sends another email, and so on. The professor replies to 5 emails every day. Compute how much time does every student wait, on average, for an answer.

Solution

Given:

$$N = 100 \qquad Z = 7 \qquad X = 5$$

Average time for a response:

$$R = \frac{N}{X} - Z = \frac{100}{5} - 7 = 13 \text{ days}$$

3.6 Exercise six

Software monitoring data for an interactive system shows a CPU utilization of 75%, a 3-second CPU service demand, a response time of 15 seconds, and 10 active users. Compute the average think time of these users.

Solution

Given:

$$U_{\text{CPU}} = 0.75 \qquad D_{\text{CPU}} = 3 \qquad R = 15 \qquad N = 10$$

We need to find Z using the response time law. Thus, we compute the throughput as:

$$X = \frac{U_{\text{CPU}}}{D_{\text{CPU}}} = \frac{0.75}{3} = 0.25$$

Now, we have:

$$Z = \frac{N}{X} - R = \frac{10}{0.25} - 15 = 25 \text{ s}$$

3.7 Exercise seven

Consider an interactive system that accommodates 100 users, each with a 15-second think time, and operates at a throughput of 5 interactions per second.

- Compute the response time of the system.
- Now, suppose the service demands of the workload shift, leading to a 50% reduction in system throughput, dropping to 2.5 interactions per second. Assume that the user count remains unchanged at 100, with the same 15-second think times, compute their response time.

Solution

Given:

$$N = 100 \qquad Z = 15 \qquad X = 5$$

1. The response time of the system can be determined using the response time law:
$$R = \frac{N}{X} - Z = \frac{100}{5} - 15 = 5 \text{ s}$$
2. If we adjust the throughput to $X' = 2.5$, the response time can be recalculated:
$$R = \frac{N}{X} - Z = \frac{100}{2.5} - 15 = 25 \text{ s}$$

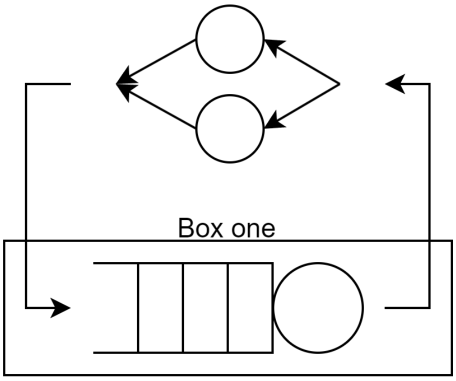
3.8 Exercise eight

A user request submitted to the system must wait in a memory queue before being processed in the central subsystem.

1. With 100 active users, each with a 20-second think time, and a system response time of 10 seconds (the sum of memory queueing and central sub-system residence times), compute how many customers are, on average, competing for memory.
2. If the memory queueing time is 8 seconds, compute the average number of customers loaded in memory.

Solution

The system is illustrated as follows:



Given:

$$N = 100 \qquad Z = 20 \qquad R = 10$$

1. Using Little's law at the memory queue, we first calculate the system throughput:

   $$X = \frac{N}{R+Z} = \frac{100}{10+20} = 3.3$$

   The number of users competing for memory:

   $$N_1 = XR = 3.3 \cdot 10 = 33.3$$

2. Given memory queueing response time $R_{mq} = 8$, the number of users waiting in memory can be computed as:

   $$N_{cs} = N_1 - N_{mq} = N_1 - N_{mq} = 33.3 - 3.3 \cdot 8 = 6.6$$

3.9 Exercise nine

Consider the following measurement data for an interactive system with a memory constraint:

- T: 1 hour.
- N: 80.
- R: 1 second.
- N in memory: 6.
- C: 36000.
- U_{CPU}: 75%.
- U_{D1}: 50%.
- U_{D2}: 50%.
- U_{D3}: 25%.

1. Compute the throughput (in requests per second).
2. Compute the average think time.
3. Compute, on the average, how many users were attempting to obtain service.
4. Compute, on the average, how much time does a user spend waiting for memory.

Solution

1. The throughput of the system is:
$$X = \frac{C}{T} = \frac{36000}{3600} = 10\,\frac{\text{req}}{\text{sec}}$$
Notice that:
$$X' = \frac{N}{R} = \frac{80}{1} = 80\,\frac{\text{req}}{\text{sec}} \neq X$$
That's because in an interactive system we have thinking users and thus we cannot use Little's law.

2. The average think time is:
$$Z = \frac{N}{X} - R = \frac{80}{10} - 1 = 7 \text{ s}$$

3. The number of users not thinking can be computed using Little's law:
$$N'' = XR = 10 \cdot 1 = 10\,\text{req}$$

4. The time spent waiting for memory is computed with Little's law:
$$R_{\text{wait}} = \frac{N_{\text{wait}}}{X} = \frac{N'' - N}{X} = \frac{10-6}{10} = 0.4 \text{ s}$$

3.10 Exercise ten

In the context of monitoring a single-class interactive system, the following data is observed:

- Disk demand: 4 seconds per transaction.
- CPU demand: 1 second per transaction.
- CPU utilization: 80%.
- Response time: 20 seconds per transaction.
- Monitoring period: 6 minutes.
- Number of users: 16.

Compute what is the average think time for these users.

Solution

Given:
$$D_d = 4 \qquad D_c = 1 \qquad U_c = 0.8 \qquad R = 20 \qquad T = 6 \qquad N = 16$$

We can calculate the average think time using the response time law, but first, we need to find the throughput:
$$X = \frac{U_c}{D_c} = \frac{0.8}{1} = 0.8\,\frac{\text{transaction}}{\text{sec}}$$

Therefore, the average think time is:
$$Z = \frac{N}{X} - R = \frac{16}{0.8} - 20 = 0 \text{ s}$$

3.11 Exercise eleven

Here are the measurement data for an interactive system:

- Measurement interval: 2 minutes.
- Number of users: 15.
- Average response time per transaction: 10 seconds.
- Disk demand: 0.5 seconds per transaction.
- CPU demand: 0.2 seconds per transaction.
- Number of servers: 20.
- Number of completed transactions: 60.

Compute the average number of users who are thinking.

Solution

We know that the total number of users is given by:

$$N = N'_{\text{think}} + N''_{\text{in}}$$

To find the number of thinking users, we first need to find the number of users in the system since N is given. We can find the system throughput as:

$$X = \frac{C}{T} = \frac{C}{T} = \frac{60}{120} = 0.5\,\frac{\text{transactions}}{\text{sec}}$$

Utilizing Little's law:

$$N'' = XR = 0.5 \cdot 10 = 5$$

Therefore, the number of users thinking is:

$$N' = N - N'' = 15 - 5 = 10$$

3.12 Exercise twelve

At a ski resort, there are 1000 visitors per day. On average, each skier visits the resort's main ski slope 10 times. Compute how many visits does the slope receive in a day.

Solution

Given:

$$x = 1000 \qquad V_{\text{slope}} = 10$$

We can find the total visits to the slope per day as:

$$X_{\text{slope}} = V_{\text{slope}} \cdot X = 10 \cdot 1000 = 10000\,\frac{\text{visits}}{\text{day}}$$

3.13 Exercise thirteen

In an interactive system with 80 active terminals, the average think time is 12 seconds per interaction. On average, each interaction results in 15 paging disk accesses. Assume that the service time per paging disk access is 30 ms and the disk is 60% busy, compute what is the average system response time.

Solution

Given:

$$N = 80 \qquad Z = 12 \qquad V_{\text{disk}} = 15 \qquad S_{\text{disk}} = 30 \qquad S_{\text{disks}} = 0.6$$

To find the average system response time, we need to determine X and X_{disk}. Using the utilization law for disks:

$$X_{\text{disk}} = \frac{U_{\text{disk}}}{S_{\text{disk}}} = \frac{0.6}{0.03} = 20\,\frac{\text{req}}{\text{sec}}$$

Then, the overall system throughput is:

$$X = \frac{X_{\text{disk}}}{V_{\text{disk}}} = \frac{20}{15} = 1.3\,\frac{\text{req}}{\text{sec}}$$

Finally, applying the response time law:

$$R = \frac{N}{X} - Z = \frac{80}{1.3} - 12 = 48 \text{ s}$$

3.14 Exercise fourteen

During a 30-minute observation interval, a specific disk was found to be busy for 12 minutes. It's known that jobs typically require 320 accesses to that disk, with an average service time per access of 25 milliseconds. Compute the system throughput (in jobs per second).

Solution

Given:

$$T = 30 \qquad B_d = 12 \qquad V_d = 320 \qquad S_d = 25$$

The throughput can be computed directly as:

$$X = \frac{U_d}{D_d} = \frac{\frac{B_d}{T}}{V_d \cdot S_d} = \frac{\frac{12}{30}}{320 \cdot 0.025} = 0.05\,\frac{\text{req}}{\text{sec}}$$

Alternatively, we have$X_d = \frac{U_d}{S_d}$, thus:

$$X = \frac{X_d}{V_d} = \frac{\frac{U_d}{S_d}}{V_d} = \frac{U_d}{V_d \cdot S_d} = 0.05\,\frac{\text{req}}{\text{sec}}$$

3.15 Exercise fifteen

In a batch system, a specific disk performs, on average, 12 operations per second. Each batch transaction requires, on average, 6 accesses to this disk. Another disk in the system handles 18 operations per second. Compute the average number of accesses to this second disk required by every batch transaction.

Solution

Given:

$$X_D = 12 \qquad V_{D1} = 6 \qquad X_{D2} = 18$$

The average number of accesses to the second disk per batch transaction is:

$$V_{D2} = \frac{X_{D2}}{X} = \frac{X_{D2}}{\frac{X_{D1}}{V_{D1}}} = \frac{18}{\frac{12}{6}} = 9$$

3.16 Exercise sixteen

The storage server of an intranet comprises two groups of disks, A and B, each with service time means $S_A = 5$ ms and $S_B = 3$ ms, respectively. The mean number of visits for the two components are $V_A = 20$ and $V_B = 30$. The throughput of A is 150 operations per second. These data were collected when the system was processing a workload generated by 300 users with a think time of $Z = 15$ s.

1. Calculate the system throughput X and the Utilization of B.
2. Determine the system response time.
3. Assume that the number of users increases to 400, compute the new response time.

Solution

1. For the system throughput, we need the utilization and the mean service time of A:

   $$U_A = X_A \cdot S_A = 0.005 \cdot 150 = 0.75$$

   $$D_A = V_A \cdot S_A = 20 \cdot 5 = 100\,\text{ms}$$

   At this point, the system throughput is:

   $$X = \frac{U_A}{D_A} = \frac{0.75}{100} = 7.5\,\frac{\text{op}}{\text{sec}}$$

   The utilization of B is:

   $$U_B = X \cdot D_B = X \cdot (S_B \cdot V_B) = D_B = 7.5 \cdot (3 \cdot 30) = 0.675$$

2. The system response time can be computed with the response time law:

   $$R = \frac{N}{X} - Z = \frac{300}{7.5} - 15 = 20\text{ s}$$

3. We cannot determine the new response time for $N = 400$ with the provided information.

3.17 Exercise seventeen

The throughput of a disk is 100 I/O operations per second. To complete a given request, 20 visits to the disk are required. There are 100 users, and the response time is 15 seconds. Compute the users' think time.

Solution

The throughput of the system is:

$$X = \frac{X_D}{V_D} = \frac{100}{20} = 5\,\frac{\text{op}}{\text{sec}}$$

Finally, the think time can be computed with the response time law:

$$Z = \frac{N}{X} - R = \frac{100}{5} - 15 = 5 \text{ s}$$

3.18 Exercise eighteen

A web server of a company is connected to an intranet and is accessed by the employees who work internally in the company, resulting in a fixed-size population of $N = 21$ users. The average think time of the users is $Z = 20$ s. A complete execution of a request generates a load of $V_s = 20$ operations to a specific storage device whose utilization is $U_s = 0.30$. The service time of the storage device per each visit is $S_s = 0.025$ s.

1. Determine the average system response time.
2. Compute the average throughput and system response time with $N = 40$ users.

Solution

1. The throughput of the system is:

   $$X = \frac{U_S}{V_S} = \frac{0.3}{20 \cdot 0.025} = 0.6\,\frac{\text{op}}{\text{sec}}$$

   Finally, the system response time can be computed with the response time law:

   $$R = \frac{N}{X} - Z = \frac{21}{0.6} - 20 = 15 \text{ s}$$

2. We cannot determine the average throughput and system response time for $N = 40$ with the provided information.

3.19 Exercise nineteen

A web application is deployed across three servers (a web + application server, and two storage servers), each characterized by the following average service times and visit rates:

- $S_{WAS} = 200$ ms, $V_{WAS} = 2$.
- $S_{SS1} = 20$ ms, $V_{SS1} = 15$.
- $S_{SS2} = 50$ ms, $V_{SS2} = 8$.

Determine:

1. The maximum throughput of the system.
2. The minimum response time.

Solution

1. We calculate the response time D for each server as:

$$\begin{cases} D_{WAS} = S_{WAS} \cdot V_{WAS} = 200 \cdot 2 = 0.4 \text{ s} \\ D_{SS1} = S_{SS1} \cdot V_{SS1} = 20 \cdot 15 = 0.3 \text{ s} \\ D_{SS2} = S_{SS2} \cdot V_{SS2} = 50 \cdot 8 = 0.4 \text{ s} \end{cases}$$

We then find the maximum response time among the three servers:

$$D_{\max} = \max\{0.4, 0.3, 0.4\} = 0.4 \text{ s}$$

The maximum throughput is the inverse of $D_{\max}$:

$$X_{\max} = \frac{1}{D_{\max}} = \frac{1}{0.4} = 2.5 \, \frac{\text{job}}{\text{sec}}$$

2. The response time of the system is the sum of the individual response times:

$$D = D_{WAS} + D_{SS1} + D_{SS2} = 0.4 + 0.3 + 0.4 = 1.1 \text{ s}$$

3.20 Exercise twenty

A batch processing system of a bank, is composed by an application server, a database server and a storage server.

- $S_{APP} = 20$ ms, $V_{APP} = 25$.
- $S_{DB} = 15$ ms, $V_{DB} = 20$.
- $S_{SS} = 50$ ms, $V_{SS} = 10$.

We are tasked to determine:

1. The maximum throughput of the system for $N = 5$.
2. The minimum response time for $N = 10$.

Solution

1. We compute the response time D for each server as:

$$\begin{cases} D_{APP} = S_{APP} \cdot V_{APP} = 20 \cdot 25 = 0.5 \text{ s} \\ D_{DB} = S_{DB} \cdot V_{DB} = 15 \cdot 20 = 0.3 \text{ s} \\ D_{SS} = S_{SS} \cdot V_{SS} = 50 \cdot 10 = 0.5 \text{ s} \end{cases}$$

We find the maximum response time among the three servers:

$$D_{\max} = \max\{0.5, 0.3, 0.5\} = 0.5 \text{ s}$$

The response time of the system is the sum of the individual response times:

$$D = D_{APP} + D_{DB} + D_{SS} = 0.5 + 0.3 + 0.5 = 1.3 \text{ s}$$

Since $N \neq 1$, the maximum throughput is computed as the minimum of $\frac{N}{D}$ and $\frac{1}{D_{\max}}$:

$$X_{\max}(5) = \min\left(\frac{N}{D}, \frac{1}{D_{\max}}\right) = \min\left(\frac{5}{1.3}, \frac{1}{0.5}\right) = \min\{3.85, 2\} = 2 \, \frac{\text{job}}{\text{sec}}$$

2. The minimum response time for $N = 10$ is:

$$R_{\min}(10) = \max\left(D, N \cdot D_{\max}\right) = \max\left(1.3, 10 \cdot 0.5\right) = \max\{1.3, 5\} = 5 \text{ s}$$

3.21 Exercise twenty-one

An interactive system, where users have a think time $Z = 200$ s, consists of an application server, a database server, and a web server.

- $S_{APP} = 20$ ms, $V_{APP} = 5$.
- $S_{DB} = 10$ ms, $V_{DB} = 20$.
- $S_{WEB} = 40$ ms, $V_{WEB} = 2$.

We need to determine:

1. The number of users for which the system switches from low to high load.
2. The minimum and maximum response time for $N = 50$.
3. The minimum and maximum throughput for $N = 2000$.

Solution

1. We compute the response time D for each server as:

$$\begin{cases} D_{APP} = S_{APP} \cdot V_{APP} = 20 \cdot 5 = 0.1 \text{ s} \\ D_{DB} = S_{DB} \cdot V_{DB} = 10 \cdot 20 = 0.2 \text{ s} \\ D_{WEB} = S_{WEB} \cdot V_{WEB} = 40 \cdot 2 = 0.08 \text{ s} \end{cases}$$

We find the maximum response time among the three servers:

$$D_{\max} = \max\{0.1, 0.2, 0.08\} = 0.2 \text{ s}$$

The response time of the system is the sum of the individual response times plus the think time:

$$D = D_{APP} + D_{DB} + D_{WEB} = 0.1 + 0.2 + 0.08 = 0.38 \text{ s}$$

The number of users for which the system switches from low to high load, N^*, is:

$$N^* = \frac{D + Z}{D_{\max}} = \frac{0.38 + 200}{0.2} = 1001.9$$

2. The minimum and maximum response time for $N = 50$ are:

$$\max\left(0.38, 50 \cdot 0.2 - 200\right) \leq R(50) \leq 50 \cdot 0.38 \rightarrow \max\{0.38, -190\} \leq R(50) \leq 19$$

Thus, the minimum response time is 0.38 seconds, while the maximum is 19 seconds.

3. The minimum and maximum throughput for $N = 2000$ are:

$$\frac{2000}{2000 \cdot 0.38 + 200} \leq X(2000) \leq \min\left(\frac{2000}{0.38 + 200}, \frac{1}{0.2}\right)$$

Thus, the minimum throughput is approximately 2.583 jobs per second, while the maximum is 5 jobs per second.

3.22 Exercise twenty-two

Let's analyze an infrastructure consisting of a web server, an application server, and a storage server. After a 1-hour measurement period, during which $N = 50$ users continuously worked, the following data were collected:

- Total number of jobs executed by the system: $C = 5400$ jobs.
- Number of WS completed operations: $C_{WS} = 54000$ operations.
- Number of AS completed operations: $C_{AS} = 32400$ operations.
- Number of SS completed operations: $C_{SS} = 10800$ operations.
- WS total activity time: $B_{WS} = 1800$ seconds.
- AS total activity time: $B_{AS} = 720$ seconds.
- SS total activity time: $B_{SS} = 900$ seconds.
- Mean think time: $Z = 5$ seconds.

Using operational analysis equations:

1. Calculate the visits V_i to the three servers during a complete job execution, their global service demands D_i, and determine the bottleneck resource of the infrastructure.
2. Compute the response time when $N = 50$ users are connected, as well as the maximum throughput when the number of users tends to infinity (asymptotic value).
3. Substitute the bottleneck resource determined in step 1 with another resource that is two times (2x) more powerful. Compute the new value of the asymptotic throughput.

Solution

1. We compute the number of visits as:

$$\begin{cases} V_{WS} = \frac{C_{WS}}{C} = \frac{54000}{5400} = 10 \text{ visits} \\ V_{AS} = \frac{C_{AS}}{C} = \frac{32400}{5400} = 6 \text{ visits} \\ V_{SS} = \frac{C_{SS}}{C} = \frac{10800}{5400} = 2 \text{ visits} \end{cases}$$

The utilization of each resource is:

$$\begin{cases} U_{WS} = \frac{B_{WS}}{T} = \frac{1800}{3600} = 0.5 \\ U_{AS} = \frac{B_{AS}}{T} = \frac{720}{3600} = 0.2 \\ U_{SS} = \frac{B_{SS}}{T} = \frac{900}{3600} = 0.25 \end{cases}$$

The throughput of the system is:

$$X = \frac{C}{T} = \frac{5400}{3600} = 1.5 \, \frac{\text{job}}{\text{second}}$$

We can now compute the global service demands as:

$$\begin{cases} D_{WS} = \frac{U_{WS}}{X} = \frac{0.5}{1.5} = 0.33 \text{ visits} \\ D_{AS} = \frac{U_{AS}}{X} = \frac{0.2}{1.5} = 0.13 \text{ visits} \\ D_{SS} = \frac{U_{SS}}{X} = \frac{0.25}{1.5} = 0.16 \text{ visits} \end{cases}$$

The maximum service demand is:

$$D_{\max} = \max\{0.33, 0.13, 0.16\} = 0.33$$

Therefore, the bottleneck is the web server.

2. The response time with $N = 50$ is:

$$R(50) = \frac{N}{X} - Z = 28.3 \text{ s}$$

The upper bound for the throughput is then:

$$X_{\max} = \frac{1}{D_{\max}} = \frac{1}{0.33} = 3 \, \frac{\text{job}}{\text{second}}$$

3. After replacing the bottlenecked web server with one that is two times more powerful, the new $D'_{\max}$ is the web server, and the bottlenecks become the web server and the storage server. The new upper bound for the throughput is then:

$$X'_{\max} = \frac{1}{D'_{\max}} = \frac{1}{0.16} = 6 \, \frac{\text{job}}{\text{second}}$$

3.23 Exercise twenty-three

Let's consider an intranet accessible by a large number of users. A single request passes through an application server with a service time $S = 300$ ms, then through a database server with a service time $S = 250$ ms, and then back through the application server. Additionally, a request must pass through the system firewall before entering and exiting the intranet, with a firewall service time per visit of $S = 10$ ms.

1. Compute the maximum throughput of the system.

2. It's possible for the response time to be less than 9 seconds. Find the conditions to let this happen.

Solution

1. We compute the response time D for each server as:

$$\begin{cases} D_{AS} = S_{AS} \cdot V_{AS} = 300 \cdot 2 = 0.6 \text{ s} \\ D_{DS} = S_{DS} \cdot V_{DS} = 250 \cdot 1 = 0.25 \text{ s} \\ D_{FW} = S_{FW} \cdot V_{FW} = 10 \cdot 2 = 0.02 \text{ s} \end{cases}$$

We find the maximum response time among the three servers:

$$D_{\text{max}} = \max\{0.6, 0.25, 0.02\} = 0.6 \text{ s}$$

The maximum throughput is the inverse of D_{max}:

$$X_{\text{max}} = \frac{1}{D_{\text{max}}} = \frac{1}{0.6} = 1.6 \,\frac{\text{job}}{\text{sec}}$$

2. For a response time $R < 9$ s, we have:

$$\max D, N \cdot D_{\text{max}} - Z \leq R(N) \leq N \cdot D$$

So, the maximum number of users N is:

$$N \cdot D \leq 9 \rightarrow N \leq 10$$

Thus, it's possible to have a response time less than 9 s if the number of users does not exceed 10.

3.24 Exercise twenty-four

Consider a session of a graphical multi-user workstation that utilizes a disk with an average service time of $S_{\text{disk}} = 25$ ms. The following measurements are obtained:

- Average think-time: $Z = 10$ s.
- Average CPU service demand: $D_{\text{CPU}} = 4$ s.
- Average disk service demand: $D_{\text{disk}} = 5$ s.
- Fraction of the busy time in which the CPU performs floating-point operations: 75%.

Using asymptotic bounds, determine which of the following modifications is more advantageous:

1. Adding a floating-point unit (FPU) that is 10 times as fast as the CPU to offload floating-point operations.
2. Replacing the disk with a new one with $S'_{\text{disk}} = 15$ ms.

Solution

The bottleneck of the system is the disk; replacing the CPU will have marginal effects. Therefore, the best solution is to replace the disk.

3.25 Exercise twenty-five

Let's examine a batch system featuring one CPU and two disks, with the following measurements:

- Monitoring period: 500 seconds.

- CPU busy time: 144 seconds.
- CPU completed operations: 200.
- Slow disk busy time: 86 seconds.
- Slow disk completed operations: 100.
- Fast disk busy time: 226 seconds.
- Fast disk completed operations: 500.
- Completed transactions: 200.
- Number of concurrent jobs: 2.

We aim to balance the load between the two disks by shifting files to increase the expected maximum throughput. After redistributing the files, what is the maximum throughput for the new system, according to asymptotic bounds? Integer values for visits are not required.

Solution

Based on the provided metrics, we derive the following calculations:

$$\begin{aligned}
D_{\text{slow}} &= \frac{B_{\text{slow}}}{C} = 0.43 \\
D_{\text{fast}} &= \frac{B_{\text{fast}}}{C} = 1.13 \\
V_{\text{fast}} &= \frac{C_{\text{fast}}}{C} = 2.5 \\
V_{\text{slow}} &= \frac{C_{\text{slow}}}{C} = 0.5 \\
V_{\text{bal}} &= V_{\text{fast}} + V_{\text{slow}} = 3 \\
S_{\text{fast}} &= \frac{D_{\text{fast}}}{\sqrt{V_{\text{fast}}}} = 0.45200 \\
S_{\text{slow}} &= \frac{D_{\text{slow}}}{\sqrt{V_{\text{slow}}}} = 0.86000 \\
D_{\text{cpu}} &= \frac{B_{\text{cpu}}}{C} = 0.72 \\
D_{\text{bal}} &= \frac{V_{\text{bal}}}{(1/S_{\text{slow}} + 1/S_{\text{fast}})} = 0.88884 \\
D_{\text{tot}} &= 2 \times D_{\text{bal}} + D_{\text{cpu}} = 2.49768 \\
D_{\text{max}_b} &= \max\left(D_{\text{bal}}, D_{\text{cpu}}\right) = 0.88884 \\
X_{\text{max}_h} &= \min\left(\frac{1}{D_{\text{maxb}}}, N \times D_{\text{tot}}\right) = 0.80074
\end{aligned}$$