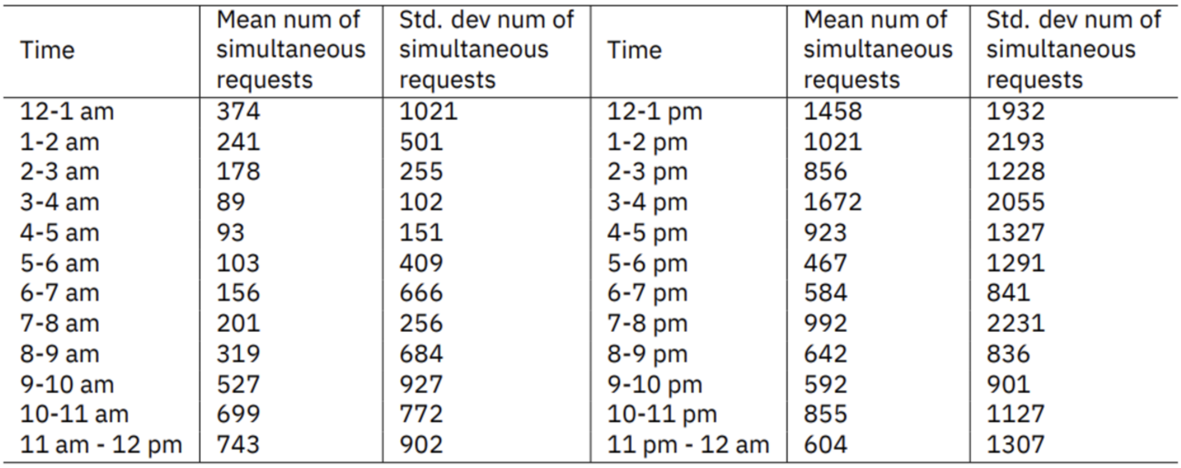
IIE3101 Stochastic Models in OR: Term Project

2022147002 김원준, 2021190002 장서현

December 14, 2024

1 Problem Description

An internet-based data center has thousands of servers to store and transmit multi-media information. One of the companies called VDO systems hosted in this data center provides digital videos to its customers at two different qualities (higher bandwidth for a higher price and vice versa). Using the prices quoted, each customer can choose which of the two bandwidths they would like for transmission (if available). Each server can handle a fixed finite amount of bandwidth (in other words, the total bandwidth requirements from all customers for both classes simultaneously downloading videos from a particular server should be less than the server's capacity).
When a server is "on", even if it is not handling any video traffic consumes a large amount of energy (for running the machine, HVAC, etc). Therefore turning a server off when it is not in use is highly desirable from an energy standpoint. However, frequently turning servers on and off reduces the lifetime of the server. Also for practical considerations assume that we need to decide how many servers should stay "on" at the beginning of every hour and that number stays a constant through the hour.
Based on historical data it is known that the number of simultaneous connections during an hour is a Pareto random variable (rounded off to the nearest integer). Assume that the connections stay on for a significant portion of an hour and so the value used for the number of simultaneous connections can be assumed to be true for the entire one-hour duration. The following table summarizes the mean and standard deviation of the number of simultaneous connections during different times of the day.



Time	Mean num of simultaneous requests	Std. dev num of simultaneous requests	Time	Mean num of simultaneous requests	Std. dev num of simultaneous requests
12-1 am	374	1021	12-1 pm	1458	1932
1-2 am	241	501	1-2 pm	1021	2193
2-3 am	178	255	2-3 pm	856	1228
3-4 am	89	102	3-4 pm	1672	2055
4-5 am	93	151	4-5 pm	923	1327
5-6 am	103	409	5-6 pm	467	1291
6-7 am	156	666	6-7 pm	584	841
7-8 am	201	256	7-8 pm	992	2231
8-9 am	319	684	8-9 pm	642	836
9-10 am	527	927	9-10 pm	592	901
10-11 am	699	772	10-11 pm	855	1127
11 am - 12 pm	743	902	11 pm - 12 am	604	1307

1.1 Assumptions

- The company has access to an infinite number of homogeneous servers, meaning no upper limit exists on the number of servers that can be activated. Yet, due to the consumption of large amount of energy when activated, only those needed to satisfy demand will be activated.
- Servers are homogeneous in terms of their total capacity but are allocated to either satisfy high quality requests or low quality requests. Each server must fully dedicate its capacity to handling one type of request during a given time period. Servers processing high-quality requests must exclusively use high-bandwidth and servers processing low-quality requests must exclusively use low-bandwidth. The prices, bandwidths, service costs, and penalty costs are different for processing different quality requests.

- Servers can be turned on or off dynamically at the start of any time period. The model assumes no delay in server activation or deactivation, enabling immediate response to demand fluctuations. Yet, the change of number of active servers between consecutive periods incurs per-cycle maintenance costs.
- The optimization model aims to maximize the net profit, defined as revenue from fulfilled requests minus operational costs, maintenance costs, and penalties. High quality streaming is priced higher to reflect greater value and penalty costs are applied to unmet requests and are higher for high quality streaming to reflect greater customer dissatisfaction.
- The total number of user requests are stochastic and the proportion of high and low quality requests within the total requests per hour are also stochastic.

1.2 Nomenclature

We first introduce the notations for sets, indices, parameters, and decision variables for the proposed optimization models as follows:

- Sets and Indices:
  - T: Set of elementary time periods (hours) $t \in T$
  - S: Set of scenarios used for representing evolution of uncertain parameters $s \in S$
- Deterministic Parameters:
  - P^H: Price for higher bandwidth
  - P^L: Price for lower bandwidth
  - Z: Total capacity of a server
  - Z^H: Capacity needed to satisfy a higher bandwidth request
  - Z^L: Capacity need to satisfy a lower bandwidth request
  - C^H: Cost of satisfying a higher bandwidth request
  - C^L: Cost of satisfying a lower bandwidth request
  - C^{on}: Cost of energy consumption of a server for an hour
  - C^m: Cost of per-cycle maintenance
  - F^H: Penalty cost for unmet higher bandwidth request
  - F^L: Penalty cost for unmet lower bandwidth request
- Stochastic Parameters:
  - $D_t(\xi)$: Total number of requests in hour t
  - $\alpha_t^H(\xi)$: Proportion of requests using HD streaming bandwidth between 0 to 1
  - $\alpha_t^L(\xi)$: Proportion of requests using SD streaming bandwidth between 0 to 1
  - ξ_s: Realization of uncertain parameters under scenario s
  - π_s: Probability of occurrence of scenario s
- Decision Variables:
  - x_t: Total number of servers to turn on for hour t
  - x_t^H: Number of high-bandwidth servers to turn on for hour t
  - x_t^L: Number of low-bandwidth servers to turn on for hour t
  - w_t: Total absolute change in the number of active servers between hour $t-1$ and t
  - w_t^H: Absolute change in the number of active high-bandwidth servers between hour $t-1$ and t
  - w_t^L: Absolute change in the number of low-bandwidth active servers between hour $t-1$ and t
  - $a_t^H(\xi)$: Number of high-quality requests satisfied in hour t
  - $a_t^L(\xi)$: Number of low-quality requests satisfied in hour t
  - $b_t^H(\xi)$: Number of high-quality requests unmet in hour t
  - $b_t^L(\xi)$: Number of low-quality requests unmet in hour t

1.3 Inputs

To solve the given problem, the following input values will be used:

- Sets and Indices:
  - $T = 24$: Optimization is performed over a 24-hour period
  - $S = 300$: Number of scenarios
- Deterministic Parameters:
  - $P^H = 15$ (USD/request): Competitive pricing for HD streaming (Netflix/Amazon Prime Video benchmarks)
  - $P^L = 10$ (USD/request): Typical SD pricing, lower than HD to attract price-sensitive customers
  - $Z = 1000$ (Mbps/server): Standard maximum capacity for enterprise-grade servers (Dell/HP)
  - $Z^H = 10$ (Mbps/request): HD streaming bandwidth requirements (Netflix/YouTube Premium benchmarks)
  - $Z^L = 2$ (Mbps/request): SD streaming bandwidth requirements, based on industry standards
  - $C^H = 1.5$ (USD/request): Costs including higher energy, bandwidth, and overhead costs for HD requests
  - $C^L = 0.5$ (USD/request): Lower energy and bandwidth costs for SD requests
  - $C^{on} = 5$ (USD/hour/server): Energy costs, including HVAC and cooling (AWS/Google data center benchmarks)
  - $C^m = 2$ (USD/server/cycle) : Cost due to wear-and-tear and maintenance from on/off switching
  - $F^H = 20$ (USD/request): Higher penalty due to customer dissatisfaction with HD service failures
  - $F^L = 10$ (USD/request): Moderate penalty for unmet SD requests
- Stochastic Parameters:
  - $D_t(\xi)$ (Requests): Total number of requests is sampled based on the Pareto Distribution
  - $\alpha_t^H(\xi) \in [0, 1]$: Proportion of requests using HD streaming bandwidth
  - $\alpha_t^L(\xi) \in [0, 1]$: Proportion of requests using SD streaming bandwidth
  - $\pi_s = 1/S$: Probability of occurrence is same for all scenarios s

1.4 Model

1.4.1 Objective Function

$$\text{Max} \quad \sum_{t \in T} \Big([-C^{on} x_t - C^m w_t] + \mathbb{E}\left[(P^H - C^H) a_t^H(\xi) + (P^L - C^L) a_t^L(\xi) - F^H b_t^H(\xi) - F^L b_t^L(\xi) \right] \Big) \quad (1)$$

The objective function (1) aims to maximize the net profit, which is defined as the total revenue minus the operational, maintenance, and penalty costs. The components of the objective function are as follows:

- Profit: The revenue is generated by fulfilling high-quality and low-quality requests. The revenue terms are $(P^H - C^H)a_t^H(\xi)$ for high-quality requests and $(P^L - C^L)a_t^L(\xi)$ for low-quality requests., where $P^H - C^H$ and $P^L - C^L$ denote the profit per request for high and low bandwidth, respectively, and $a_t^H(\xi)$ and $a_t^L(\xi)$ represent the number of fulfilled requests for each type.
- Operational Costs: The cost of keeping servers active is represented by $C^{on} x_t$, where C^{on} is the energy cost per server per hour, and x_t is the total number of servers active during hour t.
- Maintenance Costs: Maintenance costs are incurred when the number of active servers changes between consecutive time periods. This cost is expressed as $C^m w_t$, where C^m is the cost per change cycle and w_t is the absolute change in the number of servers between time $t - 1$ and t.
- Penalty Costs: Penalty costs are applied for unmet high quality and low quality requests. These terms are $F^H b_t^H(\xi)$ and $F^L b_t^L(\xi)$, where F^H and F^L denote the penalty per unmet request for high and low bandwidth, respectively, and $b_t^H(\xi)$ and $b_t^L(\xi)$ represent the number of unmet requests for each type.
- $\mathbb{E}$ represents the expectation considering the prevalent uncertainties.

1.4.2 Server Partition Constraint

$$x_t = x_t^H + x_t^L, \quad \forall t \in T \tag{2}$$

The total number of active servers during each time period is divided between servers assigned to high quality requests (x_t^H) and those assigned to low quality requests (x_t^L).

1.4.3 Transition Constraints

$$w_t = w_t^H + w_t^L, \quad \forall t \in T \tag{3a}$$
$$w_t^H \geq x_t^H - x_{t-1}^H, \quad \forall t \in T, \, t \geq 1 \tag{3b}$$
$$w_t^H \geq x_{t-1}^H - x_t^H, \quad \forall t \in T, \, t \geq 1 \tag{3c}$$
$$w_t^L \geq x_t^L - x_{t-1}^L, \quad \forall t \in T, \, t \geq 1 \tag{3d}$$
$$w_t^L \geq x_{t-1}^L - x_t^L, \quad \forall t \in T, \, t \geq 1 \tag{3e}$$

The transition constraints model the absolute changes in the number of active servers between consecutive time periods to account for maintenance costs. Constraint (3a) ensures that the total transition w_t is the sum of transitions in high quality servers (w_t^H) and low-quality servers (w_t^L). Constraints (3b) and (3c) define w_t^H as the absolute change in the number of servers handling high quality requests, while (3d) and (3e) define w_t^L as the absolute change in servers handling low quality requests.

1.4.4 Capacity Constraints

$$Z^H \cdot a_t^H(\xi) \leq Z \cdot x_t^H, \quad \forall t \in T \tag{4a}$$
$$Z^L \cdot a_t^L(\xi) \leq Z \cdot x_t^L, \quad \forall t \in T \tag{4b}$$

The capacity constraint ensures that the total bandwidth required to fulfill high-quality and low-quality requests does not exceed the total capacity of the servers assigned to each type. For high quality requests, the total bandwidth required is bounded by the total capacity of the high quality servers. For low quality requests, the total bandwidth required is bounded by the total capacity of the low quality servers.

1.4.5 Demand Constraints

$$a_t^H(\xi) + b_t^H(\xi) = \alpha_t^H(\xi) \cdot D_t(\xi), \quad \forall t \in T \tag{5a}$$
$$a_t^L(\xi) + b_t^L(\xi) = \alpha_t^L(\xi) \cdot D_t(\xi), \quad \forall t \in T \tag{5b}$$

The demand constraints ensure that the sum of fulfilled and unmet requests for both high quality and low quality streams equals the total demand in each category. These constraints reflect the relationship between the demand proportions (α_t^H and α_t^L) and the stochastic total demand (D_t) for each hour.

1.4.6 Non-negativity Constraints

$$x_t, x_t^H, x_t^L \geq 0, \quad w_t, w_t^H, w_t^L \geq 0, \quad a_t^H(\xi), a_t^L(\xi), b_t^H(\xi), b_t^L(\xi) \geq 0, \quad \forall t \in T \tag{6}$$

The non-negativity constraints enforce that all decision variables, including the number of servers, bandwidth allocations, and the number of fulfilled and unmet requests, are non-negative.

2 Solution Approach

2.1 Two-Stage Stochastic Programming Approach

The problem is formulated as a two-stage stochastic programming model to address the uncertainty in demand and the proportions of high and low quality requests.

- First Stage: The first-stage decisions include the number of servers to turn on for high-quality requests (x_t^H) and low-quality requests (x_t^L) at each time period t, as well as the absolute change in the total number of servers between consecutive periods (w_t). These decisions are made before the uncertainty is realized and must be robust against all potential realizations of the stochastic parameters in the second stage.
- Second Stage: Once the uncertain parameters, such as demand ($D_t(\xi)$) and quality proportions ($\alpha_t^H(\xi)$ and $\alpha_t^L(\xi)$), are realized for a specific scenario ξ_s, the second-stage recourse decisions are made. Recourse Decisions include allocating the bandwidth to satisfy high and low quality requests ($a_t^H(\xi), a_t^L(\xi)$) and accounting for unmet requests ($b_t^H(\xi), b_t^L(\xi)$).

Two-stage stochastic programming is chosen to balance the trade-off between operational costs, penalties, and revenue under uncertainty. First-stage decisions are made to minimize expected costs while hedging against possible future realizations of uncertainty. Recourse actions in the second stage allow the model to adapt to the realized scenarios, ensuring feasibility and cost-effectiveness.

Uncertainties within the model are discretized into a finite set of scenarios as the below:

$$\text{Max} \quad \sum_{t\in T}(-C^{on}x_t - C^m w_t) + \sum_{s\in S}\pi_s \sum_{t\in T}\Big((P^H - C^H)a_t^H(\xi_s) + (P^L - C^L)a_t^L(\xi_s) - F^H b_t^H(\xi_s) - F^L b_t^L(\xi_s)\Big) \tag{7}$$

$$\text{s.t.} \quad x_t = x_t^H + x_t^L, \quad \forall t \in T \tag{8}$$

$$w_t = w_t^H + w_t^L, \quad \forall t \in T \tag{9}$$

$$w_t^H \geq x_t^H - x_{t-1}^H, \quad \forall t \in T,\ t \geq 1 \tag{10}$$

$$w_t^H \geq x_{t-1}^H - x_t^H, \quad \forall t \in T,\ t \geq 1 \tag{11}$$

$$w_t^L \geq x_t^L - x_{t-1}^L, \quad \forall t \in T,\ t \geq 1 \tag{12}$$

$$w_t^L \geq x_{t-1}^L - x_t^L, \quad \forall t \in T,\ t \geq 1 \tag{13}$$

$$Z^H \cdot a_t^H(\xi_s) \leq Z \cdot x_t^H, \quad \forall t \in T, \forall s \in S \tag{14}$$

$$Z^L \cdot a_t^L(\xi_s) \leq Z \cdot x_t^L, \quad \forall t \in T, \forall s \in S \tag{15}$$

$$a_t^H(\xi_s) + b_t^H(\xi_s) = \alpha_t^H(\xi_s) \cdot D_t(\xi_s), \quad \forall t \in T, \forall s \in S \tag{16}$$

$$a_t^L(\xi_s) + b_t^L(\xi_s) = \alpha_t^L(\xi_s) \cdot D_t(\xi_s), \quad \forall t \in T, \forall s \in S \tag{17}$$

$$\sum_{s\in S}\pi_s = 1 \tag{18}$$

$$x_t, x_t^H, x_t^L \geq 0, \quad w_t, w_t^H, w_t^L \geq 0, \quad \forall t \in T \tag{19}$$

$$a_t^H(\xi_s) \geq 0, \quad a_t^L(\xi_s) \geq 0, \quad b_t^H(\xi_s) \geq 0, \quad b_t^L(\xi_s) \geq 0, \quad \forall t \in T,\ \forall s \in S \tag{20}$$

2.2 Sampling-Based Approach

A sampling-based approach is used to realize the uncertainty inherent in some of the distributions that will be used to solve this problem. The random variable distributions corresponding to the number of simultaneous requests and the proportion of high-quality requests and low-quality requests for each time period are not fixed and able to change depending on the situation, unlike other values. Therefore, these distributions should be expressed as discrete set of scenarios with the same probability of occurrence.

Using Monte Carlo simulation, the number of simultaneous requests is sampled from the Pareto distribution using inverse transform sampling for each time period. The proportion of high-quality requests and low-quality requests are generated using a combination of the cosine function and random fluctuations. At night, people often need high-quality transmissions, such as watching videos at home. On the other hand, during the daytime, relatively low-quality transmissions takes place such as sending emails at work. With this in mind, we could generate a request ratio using the waveform of a cosine function with midnight and noon as the peaks and troughs of high-quality requests respectively. We assumed a 60% occupancy rate at the peak and a 40% occupancy rate at the trough. On top of this base distribution, we've added some noise to give it some randomness. The noise has a max of 0.1 and a min of -0.1, and the fluctuation is generated using a uniform distribution.

Then, the stochastic parameters and variables become scenario-dependent parameters and variables indexed as $\xi_s\ \forall s \in S$. Based on this sampling-based approach, our main problem can be converted and formulated as a mixed-integer linear program. Reformulated problems can be solved using an off-the shelf optimization solver *Gurobi* for the set of scenarios representing the realization of stochastic parameters.

3 Computational Experiment

Given α and minimum x_m of the Pareto distribution, the mean μ and variance σ^2 are defined as follows.

$$\mu = \frac{\alpha x_m}{\alpha - 1}, \quad \sigma^2 = \frac{\alpha x_m^2}{(\alpha - 1)(\alpha - 2)} \tag{21}$$

If we organize the relationship between these two values, we can obtain the following conclusion.

$$\frac{\sigma^2}{\mu^2} = \frac{\alpha - 1}{\alpha(\alpha - 2)}, \quad \alpha > 2 \tag{22}$$

Therefore, since we know the mean and variance, we can find α of the Pareto distribution. If we plug α we found in this way back into the mean equation at (21), we can also find the minimum x_m. α and x_m can be expressed as formulas as follows.

$$\alpha = \frac{2\sigma^2 + \mu^2 + \sqrt{4\sigma^4 + \mu^4}}{2\sigma^2} \tag{23}$$

$$x_m = \frac{\mu(\alpha - 1)}{\alpha} \tag{24}$$

The cumulative distribution function of the Pareto distribution is defined as follows.

$$f(x) = 1 - \left(\frac{x_m}{x}\right)^\alpha \tag{25}$$

The process of finding the inverse function is as follows.

$$x = 1 - \left(\frac{x_m}{y}\right)^\alpha \tag{26a}$$

$$\left(\frac{x_m}{y}\right)^\alpha = 1 - x \tag{26b}$$

$$\alpha \ln\left(\frac{x_m}{y}\right) = \ln(1 - x) \tag{26c}$$

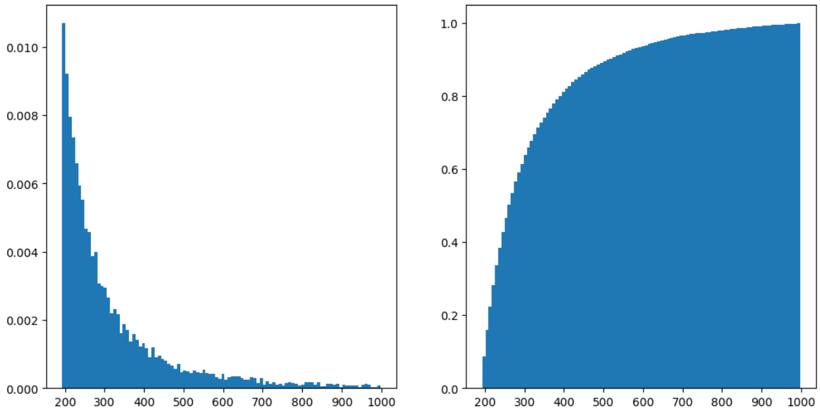
$$\ln x_m - \ln y = \frac{1}{\alpha}\ln(1 - x) \tag{26d}$$

$$\ln y = \ln x_m - \frac{1}{\alpha}\ln(1 - x) \tag{26e}$$

$$\ln y = \ln \frac{x_m}{(1 - x)^{\frac{1}{\alpha}}} \tag{26f}$$

$$y = \frac{x_m}{(1 - x)^{\frac{1}{\alpha}}} \tag{26g}$$

The probability density function and cumulative density function of the Monte Carlo simulation using the mean and standard deviation values of 12-1 am through the derived inverse function are as shown in the following graph.



To implement the distribution of quality proportion defined above, the following equation was used.

$$a_t^H(\xi) = 0.5 + \cos \frac{t}{12}\pi + U\left[-0.1, 0.1\right] \tag{27}$$

$$a_t^L(\xi) = 1 - a_t^H(\xi) \tag{28}$$

$U\left[-0.1, 0.1\right]$ is a uniform random variable from -0.1 to 0.1.
We obtained the values multiple times by applying different random seeds through 100, 300, 500, 1000, and 1500 trials, and calculated the 95% confidence interval assuming a t-distribution. The results of finding the optimal solution for the model using these distributions are as follows.

Table 1: Results of x_t

Time	**Number of trials**				
	100	300	500	1000	1500
12-1 am	20.8233±3.2397	18.0667±1.4844	19.9286±1.4755	20.9115±1.2788	20.8770±1.0926
1-2 am	11.7549±1.5989	12.2093±1.2046	12.8485±1.1546	12.7642±0.7962	12.5013±0.7578
2-3 am	7.5249±0.9784	7.5333±0.5583	7.5563±0.4866	7.7472±0.3997	7.8200±0.3396
3-4 am	3.3950±0.4168	3.2325±0.2205	3.2575±0.2239	3.3382±0.1678	3.4121±0.1718
4-5 am	4.4158±0.8795	4.3276±0.4602	4.1514±0.3871	4.3378±0.3158	4.3067±0.2513
5-6 am	5.2769±1.0838	4.9799±0.4877	4.9626±0.4490	4.9552±0.2882	5.1695±0.2511
6-7 am	7.3938±1.0895	7.6294±0.8395	7.5462±0.6581	7.8390±0.4801	7.8536±0.4347
7-8 am	7.7807±0.9990	7.1375±0.5518	7.4963±0.6584	7.4809±0.4298	7.4764±0.3451
8-9 am	16.6156±4.8939	15.0835±2.1037	14.1206±1.3729	14.2646±1.0389	14.6854±1.1867
9-10 am	20.4830±4.6258	20.5632±2.0544	22.0689±2.0347	21.0301±1.3026	21.7897±1.3091
10-11 am	24.3610±6.5615	21.7314±2.6784	21.9155±1.6626	21.9458±1.3428	21.9668±1.0749
11-12 pm	25.4497±5.4044	25.0890±2.5383	23.8798±1.7028	23.2476±1.0955	23.1607±0.9960
12-1 pm	51.4548±6.9120	49.6660±3.9445	51.8671±5.1644	50.7837±3.1670	49.2108±2.3000
1-2 pm	40.1065±4.4479	41.8316±5.1704	42.3031±3.9259	43.0565±2.7703	53.6065±1.2496
2-3 pm	29.0120±3.2413	32.9944±4.5360	31.0084±2.6461	30.7166±1.7451	30.1489±1.3630
3-4 pm	54.3982±11.4076	51.6966±5.1381	53.7913±4.2799	55.5891±3.0377	58.2069±2.6929
4-5 pm	32.6441±8.7378	34.0702±4.2287	34.9491±3.8726	35.7903±3.0365	35.3344±2.2990
5-6 pm	24.1466±3.7950	22.7628±2.5237	23.6804±2.3042	23.0772±1.6314	22.3810±1.2463
6-7 pm	22.8606±2.9878	22.6687±2.2310	23.7721±2.1065	23.4620±1.6095	23.4104±1.2340
7-8 pm	44.0997±8.9097	46.7185±5.1032	49.3358±5.4355	48.6705±3.4983	47.5121±2.5630
8-9 pm	22.4217±2.4284	23.4812±1.4535	26.3819±3.2171	25.6054±1.8934	25.9651±1.4136
9-10 pm	27.8950±3.7025	27.7542±2.0616	26.7666±2.0119	27.8481±1.5761	27.5032±1.3169
10-11 pm	37.0975±4.8228	35.2555±3.1244	35.5663±2.5141	36.1510±2.1541	36.0178±1.7773
11-12 am	35.5461±9.3168	32.3859±4.1531	31.2585±2.6846	31.8721±3.2462	31.7202±2.7587

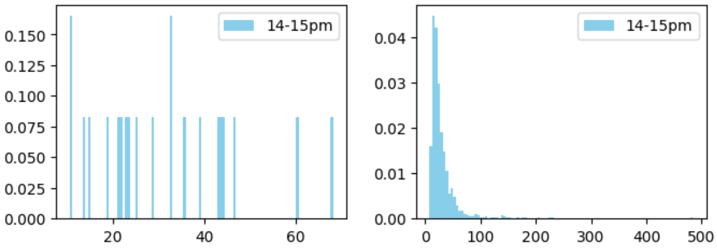
The proportion of high quality requests in each time zone is as follows (unit: %):

Table 2: Results of High Quality Requests

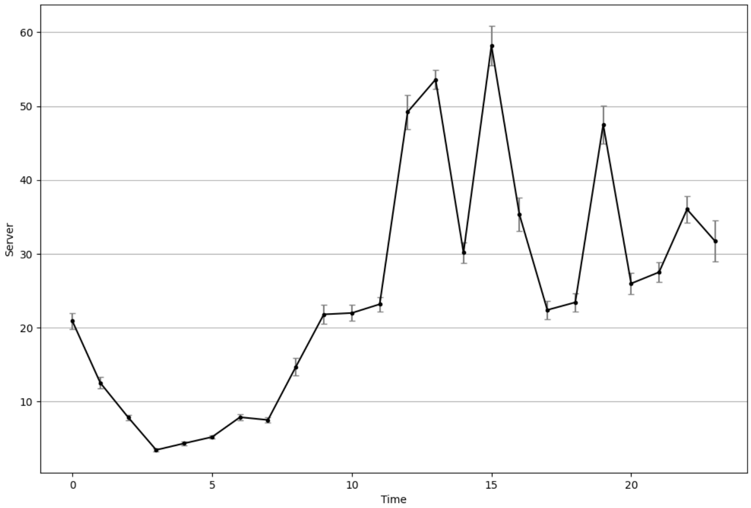
Time	**Number of trials**				
	100	300	500	1000	1500
12-1 am	88.0299±0.4831	87.9786±0.2714	87.9549±0.2121	87.9619±0.1470	87.8845±0.1166
1-2 am	87.9033±0.4244	87.7999±0.2544	87.8474±0.2013	87.8331±0.1390	87.8396±0.1147
2-3 am	87.2923±0.4649	87.1925±0.2601	87.3339±0.2015	87.3804±0.1460	87.3357±0.1192
3-4 am	87.0251±0.4760	86.8264±0.2759	86.7829±0.2160	86.6899±0.1509	86.7079±0.1217
4-5 am	85.9398±0.5068	85.9830±0.2966	85.8663±0.2264	85.7738±0.1585	85.7648±0.1294
5-6 am	84.1761±0.5734	84.2346±0.3173	84.2518±0.2515	84.3557±0.1759	84.3752±0.1433
6-7 am	82.7753±0.5334	83.0361±0.3317	83.2265±0.2560	83.2066±0.1846	83.1747±0.1493
7-8 am	81.9665±0.6406	81.7829±0.3633	81.6958±0.2830	81.7392±0.1984	81.6993±0.1605
8-9 am	80.0521±0.6678	80.1895±0.3878	80.0994±0.2955	80.1031±0.2048	80.0721±0.1671
9-10 am	78.7636±0.7221	78.8999±0.4000	78.9865±0.3127	78.9332±0.2189	78.8993±0.1776
10-11 am	77.6249±0.7239	77.7773±0.4204	77.6944±0.3252	77.8084±0.2311	77.8186±0.1896

Time	Number of trials				
	100	300	500	1000	1500
11-12 pm	77.2816±0.7041	77.0406±0.4219	76.9968±0.3260	77.0981±0.2346	77.0971±0.1907
12-1 pm	76.9972±0.8040	76.7706±0.4717	76.8822±0.3552	76.9370±0.2425	76.8863±0.1972
1-2 pm	76.9343±0.7679	77.0867±0.4327	76.9771±0.3456	77.0865±0.2409	77.0512±0.1993
2-3 pm	77.5613±0.7512	77.8227±0.4180	77.8747±0.3158	77.8230±0.2271	77.7452±0.1866
3-4 pm	79.1636±0.6162	79.0499±0.3822	78.8539±0.3024	78.9162±0.2190	78.8798±0.1795
4-5 pm	80.4900±0.6233	80.5359±0.3654	80.5295±0.2861	80.2741±0.2069	80.2546±0.1698
5-6 pm	81.2621±0.5841	81.4878±0.3535	81.6122±0.2754	81.6502±0.1952	81.6002±0.1594
6-7 pm	82.7777±0.6103	83.0139±0.3319	83.0950±0.2619	83.1063±0.1848	83.1227±0.1508
7-8 pm	84.5040±0.5193	84.4782±0.3082	84.4155±0.2468	84.4318±0.1728	84.4765±0.1412
8-9 pm	85.8302±0.4927	85.8406±0.2895	85.7881±0.2231	85.8022±0.1577	85.7534±0.1304
9-10 pm	86.8135±0.4684	86.8721±0.2714	86.8643±0.2120	86.7943±0.1507	86.7508±0.1225
10-11 pm	87.5742±0.4364	87.4452±0.2529	87.4583±0.2015	87.4361±0.1421	87.4312±0.1173
11-12 pm	87.7726±0.4353	87.9884±0.2634	88.0017±0.2051	87.8807±0.1465	87.8163±0.1181

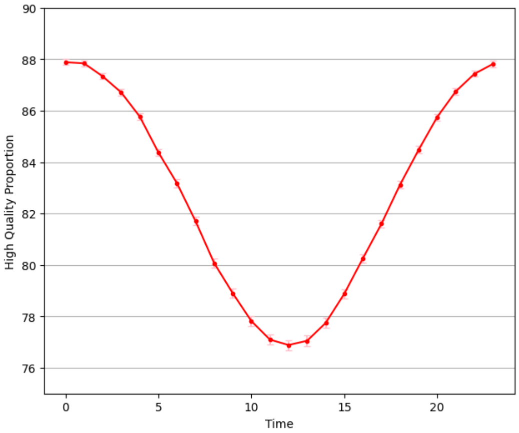
This proportion is also affected by the previously defined quality preference ratio of people by time zone, but the ratio of high quality may appear to increase more because higher quality transmissions inherently require more capacity. The proportion of low quality requests can be calculated by subtracting those values from 100 respectively. Comparing the distribution of results for 100 trials (left) and 1500 trials (right), we can see that the one with a larger number of trials is more likely to be assumed to be a normal distribution.

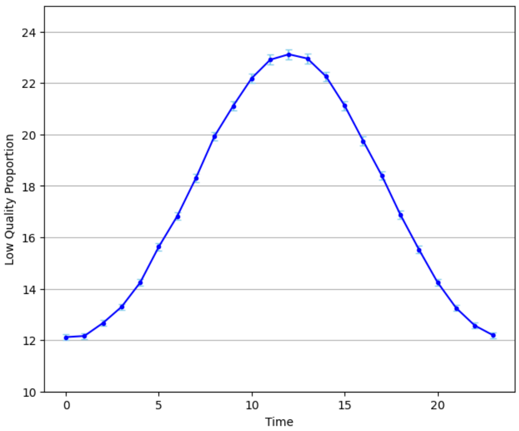


Therefore, we decided to use the results obtained through 1500 trials as the final result. The results of x_t are displayed in a graph as follows.



The black dots represent the average number of servers that should be turned on during that time period, and the gray lines represent the 95% confidence intervals. The graph on the next page shows how servers should be split to handle high and low quality requests for each time zone.





The graph on the left shows the proportion of servers that respond to high-quality requests, while the graph on the right shows the proportion of servers that respond to low-quality requests. As with the previous graph, the dots represent the mean for the respective time zone, and the vertical line represents the 95% confidence interval. These results tend to follow the quality preference ratio we defined earlier, but it is noteworthy that during the night time when there are the most high-quality requests, the proportion is split at the peak of about 88%:12%, while during the day when there are the fewest, the proportion is split at the trough of about 77%:23%.

4 Conclusion

(a) How many servers should be on during each hour of the day? (Values of x_t)
(b) How should the available bandwidth during an hour be partitioned into high and low-quality videos? (Values of x_t^H and x_t^L with their proportions)

Table 3: Final Solution

Time	(a)	(b)	
	x_t	x_t^H	x_t^L
12-1 am	20.8770	18.3476 (87.8845%)	2.5294 (12.1155%)
1-2 am	12.5013	10.9811 (87.8396%)	1.5202 (12.1604%)
2-3 am	7.8200	6.8297 (87.3357%)	0.9903 (12.6643%)
3-4 am	3.4121	2.9586 (86.7079%)	0.4535 (13.2921%)
4-5 am	4.3067	3.6936 (85.7648%)	0.6131 (14.2352%)
5-6 am	5.1695	4.3618 (84.3752%)	0.8077 (15.6248%)
6-7 am	7.8536	6.5322 (83.1747%)	1.3214 (16.8253%)
7-8 am	7.4764	6.1082 (81.6993%)	1.3682 (18.3007%)
8-9 am	14.6854	11.7589 (80.0721%)	2.9265 (19.9279%)
9-10 am	21.7897	17.1920 (78.8993%)	4.5978 (21.1007%)
10-11 am	21.9668	17.0942 (77.8186%)	4.8726 (22.1814%)
11-12 pm	23.1607	17.8562 (77.0971%)	5.3045 (22.9029%)
12-1 pm	49.2108	37.8363 (76.8863%)	11.3744 (23.1137%)
1-2 pm	53.6065	41.3044 (77.0512%)	12.3021 (22.9488%)
2-3 pm	30.1489	23.4393 (77.7452%)	6.7096 (22.2548%)
3-4 pm	58.2069	45.9135 (78.8798%)	12.2934 (21.1202%)
4-5 pm	35.3344	28.3575 (80.2546%)	6.9769 (19.7454%)
5-6 pm	22.3810	18.2629 (81.6002%)	4.1181 (18.3998%)
6-7 pm	23.4104	19.4594 (83.1227%)	3.9511 (16.8773%)
7-8 pm	47.5121	40.1366 (84.4765%)	7.3756 (15.5235%)
8-9 pm	25.9651	22.2659 (85.7534%)	3.6991 (14.2466%)
9-10 pm	27.5032	23.8593 (86.7508%)	3.6440 (13.2492%)
10-11 pm	36.0178	31.4908 (87.4312%)	4.5270 (12.5688%)
11-12 am	31.7202	27.8555 (87.8163%)	3.8647 (12.1837%)