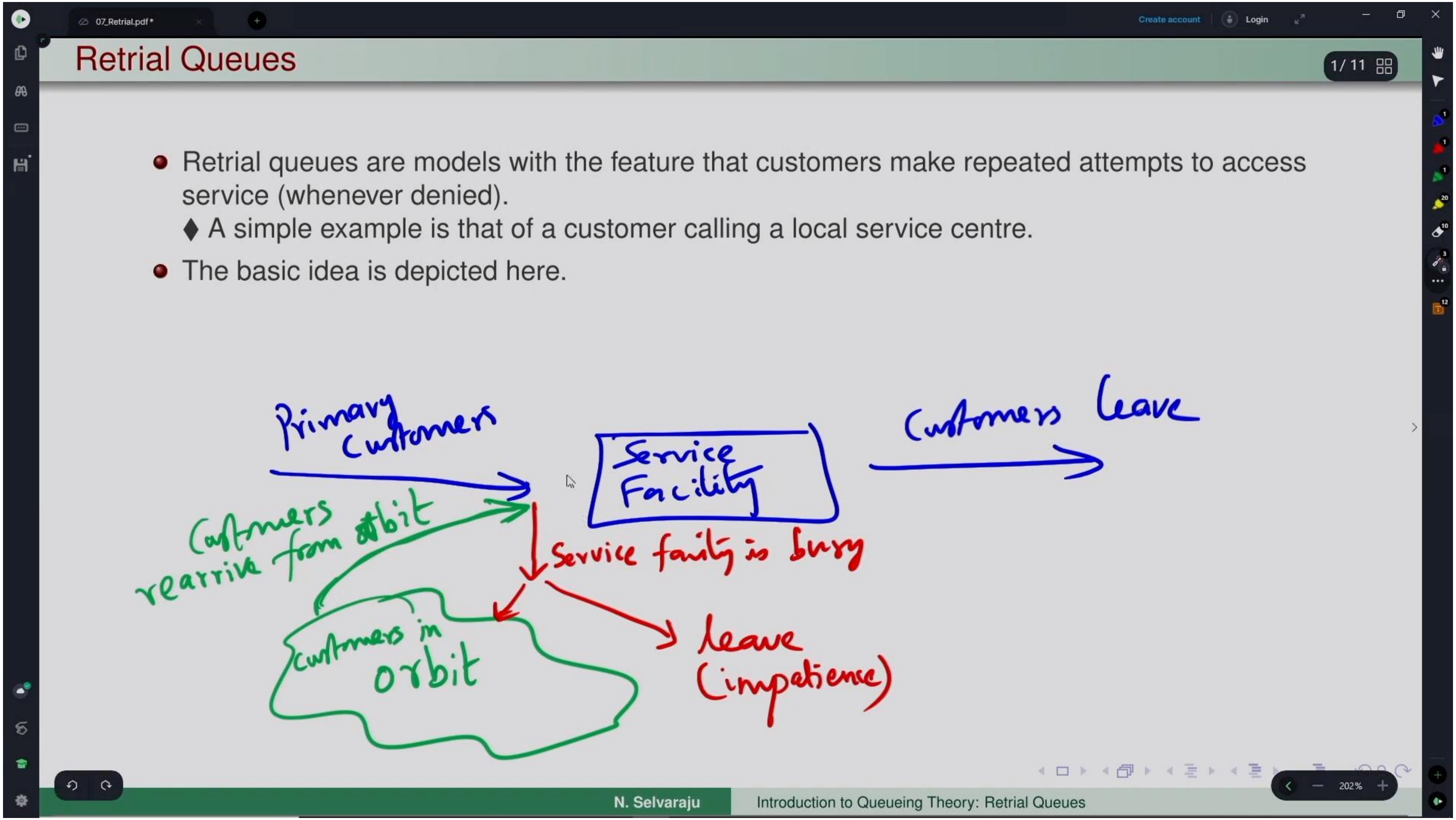
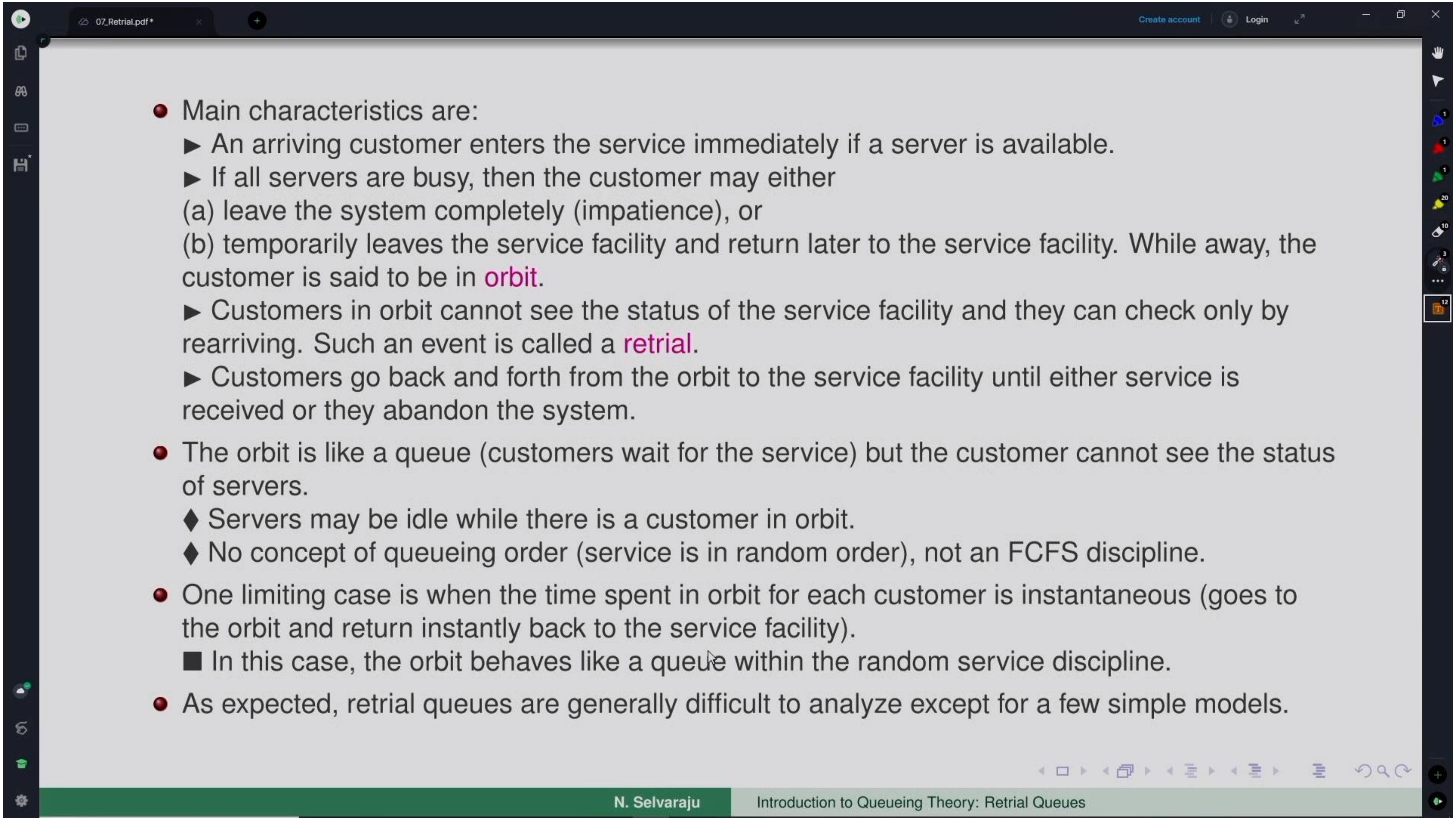
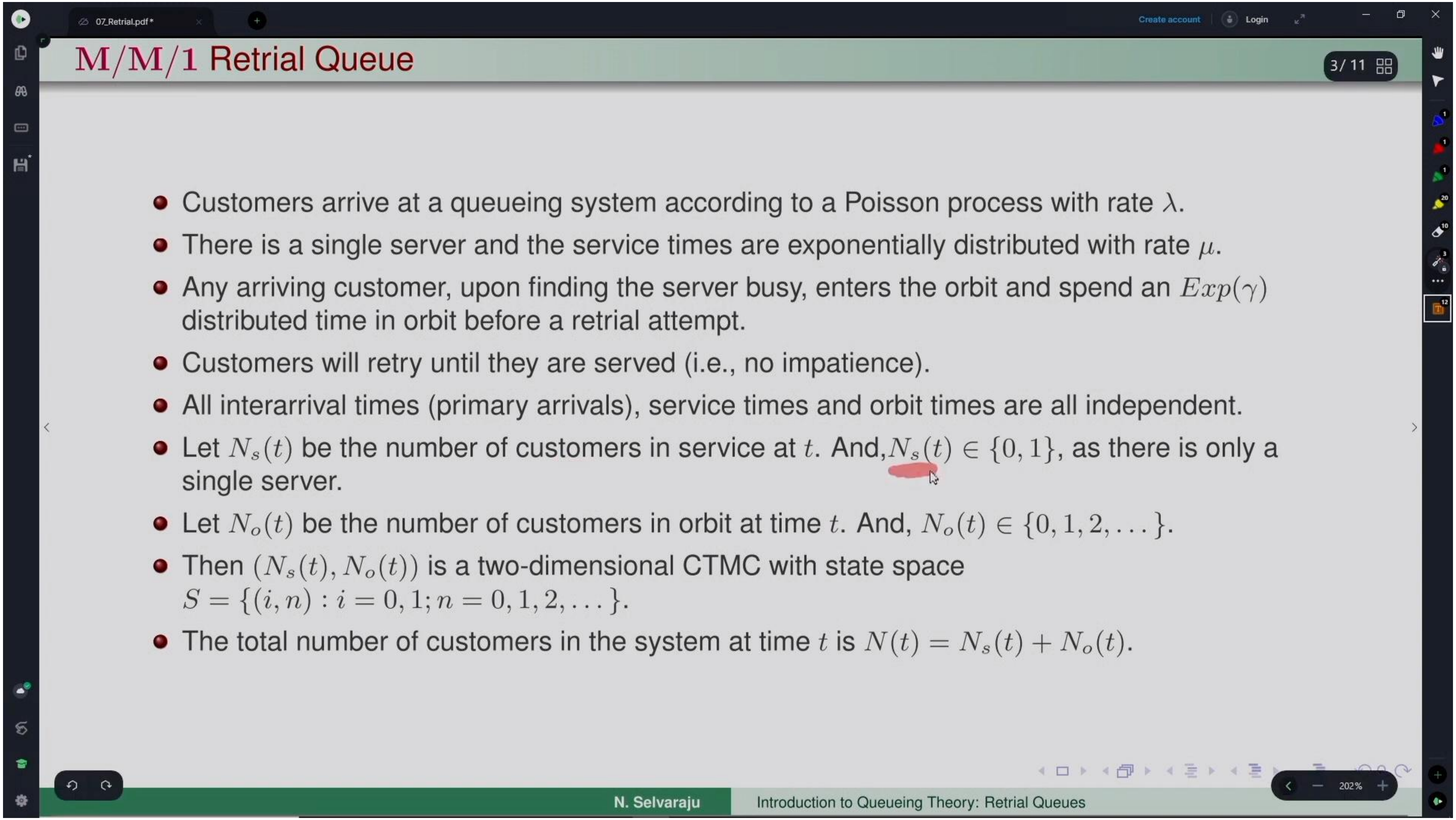


Retrial Queues

- Retrial queues are models with the feature that customers make repeated attempts to access service (whenever denied).
  - ♦ A simple example is that of a customer calling a local service centre.
- The basic idea is depicted here.

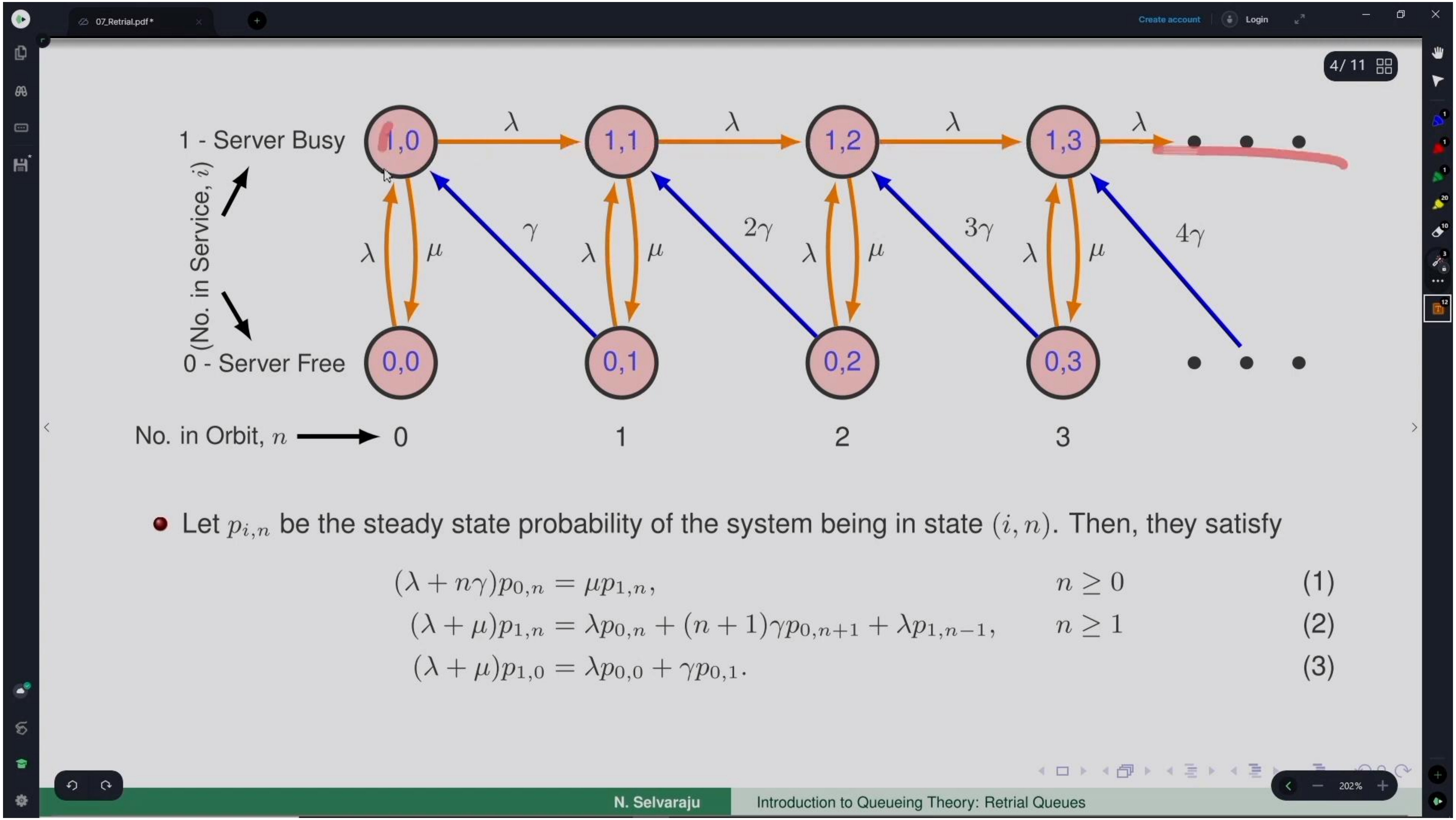


- Main characteristics are:
  ▶ An arriving customer enters the service immediately if a server is available.
  ▶ If all servers are busy, then the customer may either
  (a) leave the system completely (impatience), or
  (b) temporarily leaves the service facility and return later to the service facility. While away, the customer is said to be in orbit.
  ▶ Customers in orbit cannot see the status of the service facility and they can check only by rearriving. Such an event is called a retrial.
  ▶ Customers go back and forth from the orbit to the service facility until either service is received or they abandon the system.
- The orbit is like a queue (customers wait for the service) but the customer cannot see the status of servers.
  ⧫ Servers may be idle while there is a customer in orbit.
  ⧫ No concept of queueing order (service is in random order), not an FCFS discipline.
- One limiting case is when the time spent in orbit for each customer is instantaneous (goes to the orbit and return instantly back to the service facility).
  ■ In this case, the orbit behaves like a queue within the random service discipline.
- As expected, retrial queues are generally difficult to analyze except for a few simple models.



M/M/1 Retrial Queue

- Customers arrive at a queueing system according to a Poisson process with rate λ.
- There is a single server and the service times are exponentially distributed with rate μ.
- Any arriving customer, upon finding the server busy, enters the orbit and spend an $Exp(\gamma)$ distributed time in orbit before a retrial attempt.
- Customers will retry until they are served (i.e., no impatience).
- All interarrival times (primary arrivals), service times and orbit times are all independent.
- Let $N_s(t)$ be the number of customers in service at t. And,$N_s(t) \in \{0, 1\}$, as there is only a single server.
- Let $N_o(t)$ be the number of customers in orbit at time t. And, $N_o(t) \in \{0, 1, 2, \ldots\}$.
- Then $(N_s(t), N_o(t))$ is a two-dimensional CTMC with state space $S = \{(i, n) : i = 0, 1; n = 0, 1, 2, \ldots\}$.
- The total number of customers in the system at time t is $N(t) = N_s(t) + N_o(t)$.

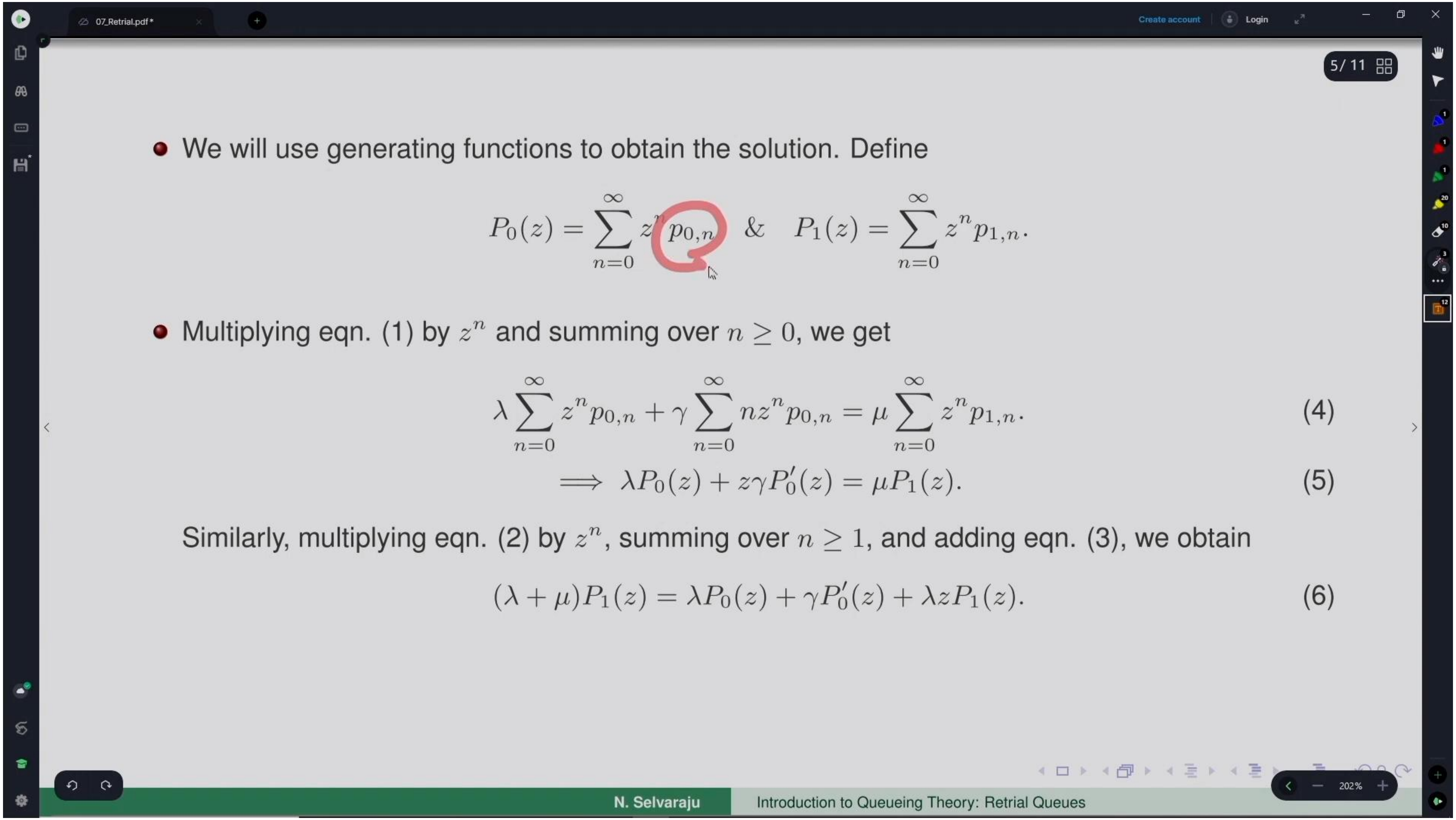


- Let $p_{i,n}$ be the steady state probability of the system being in state (i,n). Then, they satisfy

$$(\lambda + n\gamma)p_{0,n} = \mu p_{1,n}, \qquad n \geq 0 \tag{1}$$

$$(\lambda + \mu)p_{1,n} = \lambda p_{0,n} + (n+1)\gamma p_{0,n+1} + \lambda p_{1,n-1}, \qquad n \geq 1 \tag{2}$$

$$(\lambda + \mu)p_{1,0} = \lambda p_{0,0} + \gamma p_{0,1}. \tag{3}$$



- We will use generating functions to obtain the solution. Define

$$P_0(z) = \sum_{n=0}^{\infty} z^n p_{0,n} \quad \& \quad P_1(z) = \sum_{n=0}^{\infty} z^n p_{1,n}.$$

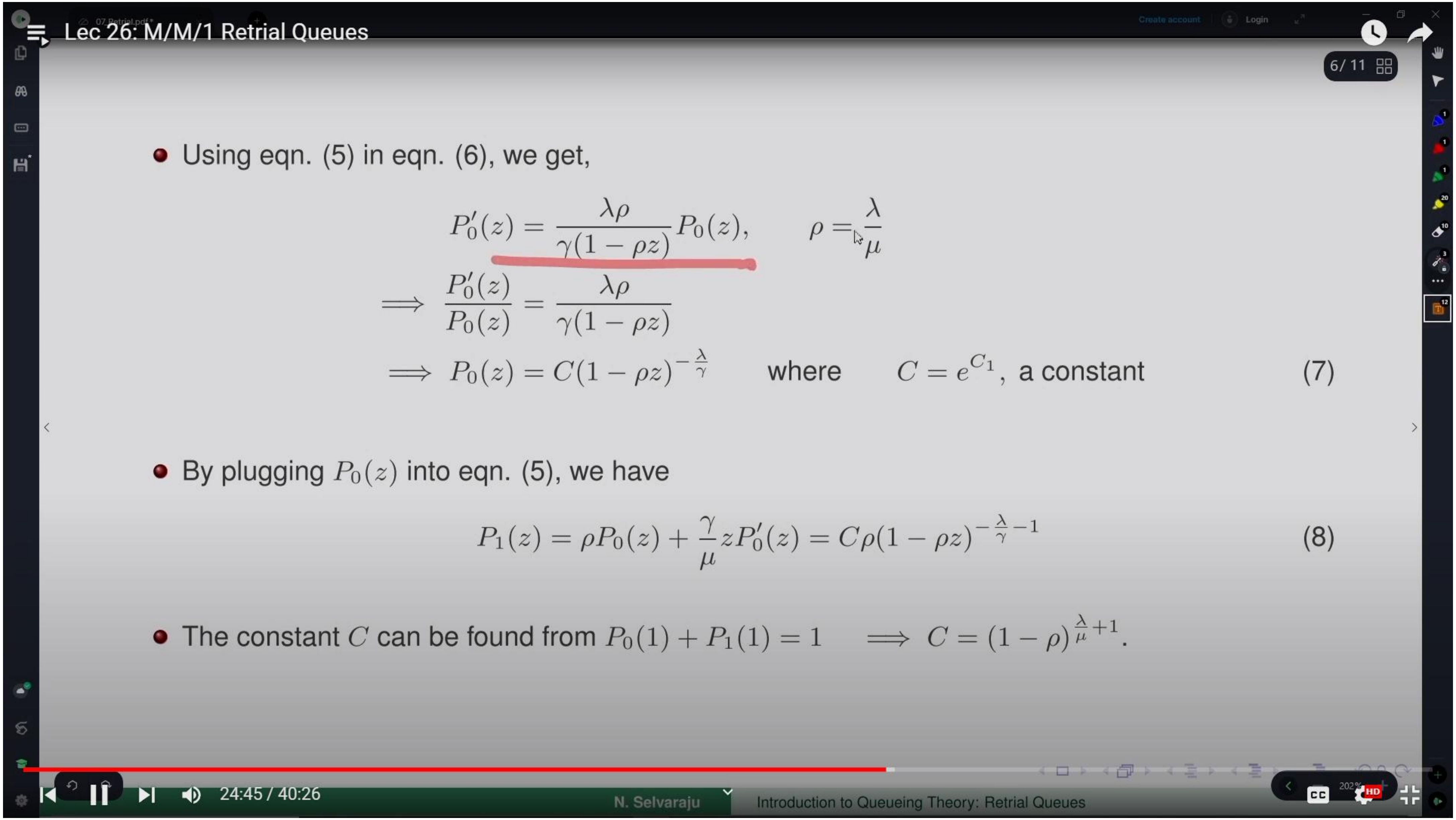
- Multiplying eqn. (1) by z^n and summing over $n \geq 0$, we get

$$\lambda \sum_{n=0}^{\infty} z^n p_{0,n} + \gamma \sum_{n=0}^{\infty} n z^n p_{0,n} = \mu \sum_{n=0}^{\infty} z^n p_{1,n}. \tag{4}$$

$$\implies \lambda P_0(z) + z\gamma P_0'(z) = \mu P_1(z). \tag{5}$$

Similarly, multiplying eqn. (2) by z^n, summing over $n \geq 1$, and adding eqn. (3), we obtain

$$(\lambda + \mu) P_1(z) = \lambda P_0(z) + \gamma P_0'(z) + \lambda z P_1(z). \tag{6}$$



Lec 26: M/M/1 Retrial Queues

- Using eqn. (5) in eqn. (6), we get,

$$P_0'(z) = \underline{\frac{\lambda\rho}{\gamma(1-\rho z)} P_0(z),} \qquad \rho = \frac{\lambda}{\mu}$$

$$\Longrightarrow \frac{P_0'(z)}{P_0(z)} = \frac{\lambda\rho}{\gamma(1-\rho z)}$$

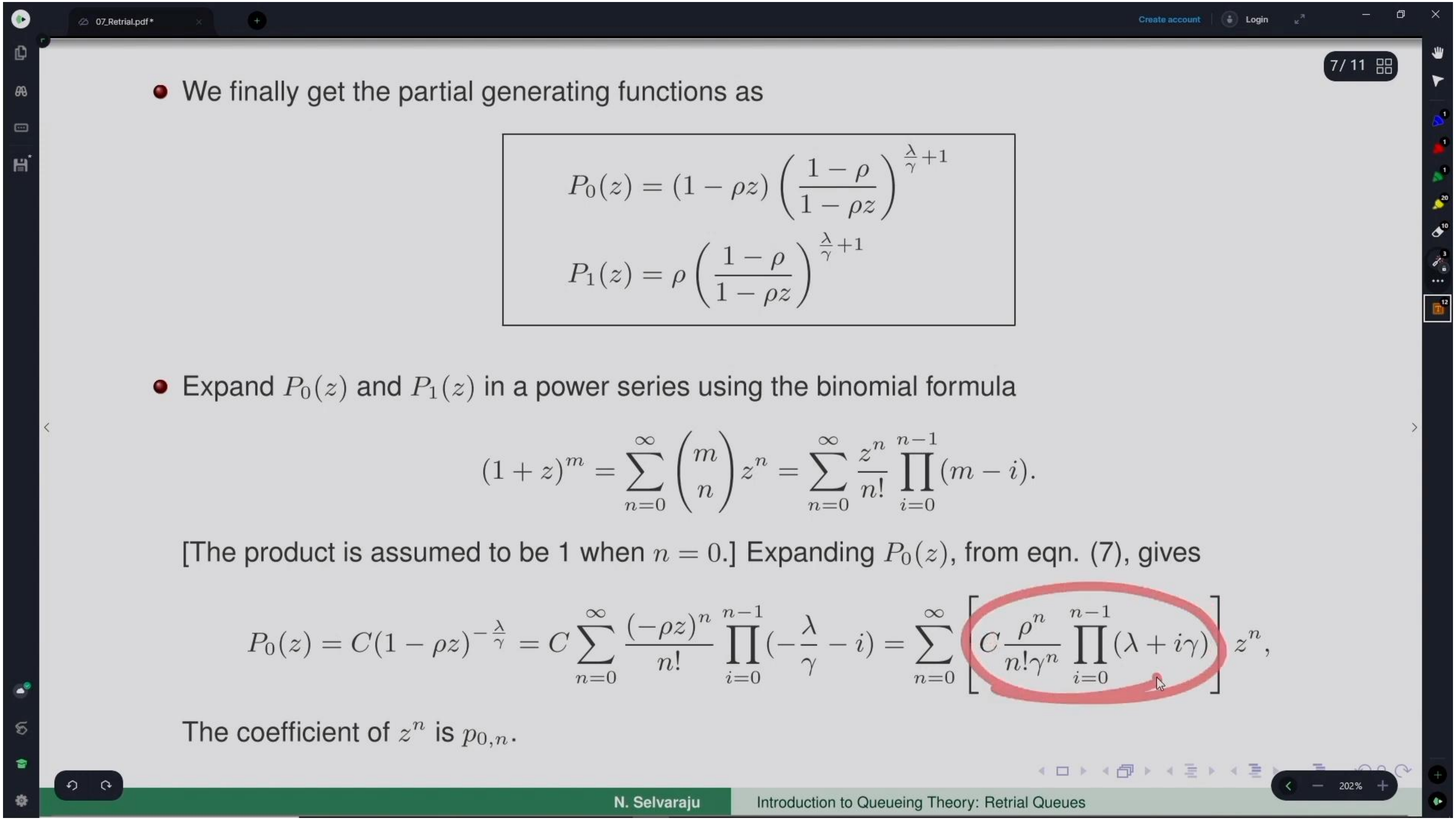
$$\Longrightarrow P_0(z) = C(1-\rho z)^{-\frac{\lambda}{\gamma}} \qquad \text{where} \qquad C = e^{C_1}, \text{ a constant} \qquad (7)$$

- By plugging $P_0(z)$ into eqn. (5), we have

$$P_1(z) = \rho P_0(z) + \frac{\gamma}{\mu} z P_0'(z) = C\rho(1-\rho z)^{-\frac{\lambda}{\gamma}-1} \qquad (8)$$

- The constant C can be found from $P_0(1) + P_1(1) = 1 \quad \Longrightarrow C = (1-\rho)^{\frac{\lambda}{\mu}+1}$.

Introduction to Queueing Theory: Retrial Queues



- We finally get the partial generating functions as

$$\boxed{\begin{aligned} P_0(z) &= (1-\rho z)\left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\lambda}{\gamma}+1} \\ P_1(z) &= \rho\left(\frac{1-\rho}{1-\rho z}\right)^{\frac{\lambda}{\gamma}+1} \end{aligned}}$$

- Expand $P_0(z)$ and $P_1(z)$ in a power series using the binomial formula

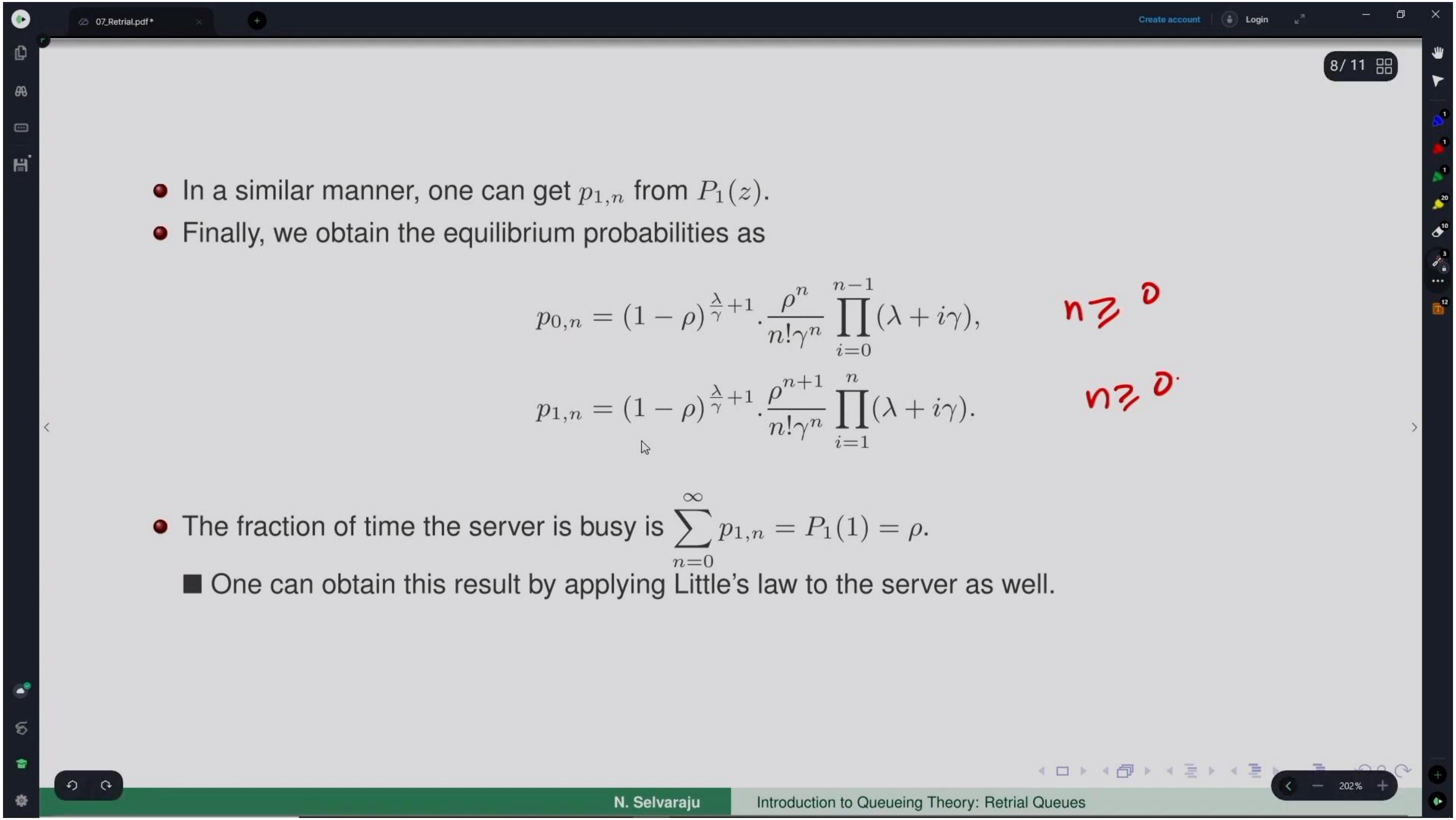
$$(1+z)^m = \sum_{n=0}^{\infty} \binom{m}{n} z^n = \sum_{n=0}^{\infty} \frac{z^n}{n!} \prod_{i=0}^{n-1} (m-i).$$

[The product is assumed to be 1 when $n = 0$.] Expanding $P_0(z)$, from eqn. (7), gives

$$P_0(z) = C(1-\rho z)^{-\frac{\lambda}{\gamma}} = C \sum_{n=0}^{\infty} \frac{(-\rho z)^n}{n!} \prod_{i=0}^{n-1} \left(-\frac{\lambda}{\gamma} - i\right) = \sum_{n=0}^{\infty} \left[C \frac{\rho^n}{n!\gamma^n} \prod_{i=0}^{n-1} (\lambda + i\gamma) \right] z^n,$$

The coefficient of z^n is $p_{0,n}$.

N. Selvaraju Introduction to Queueing Theory: Retrial Queues

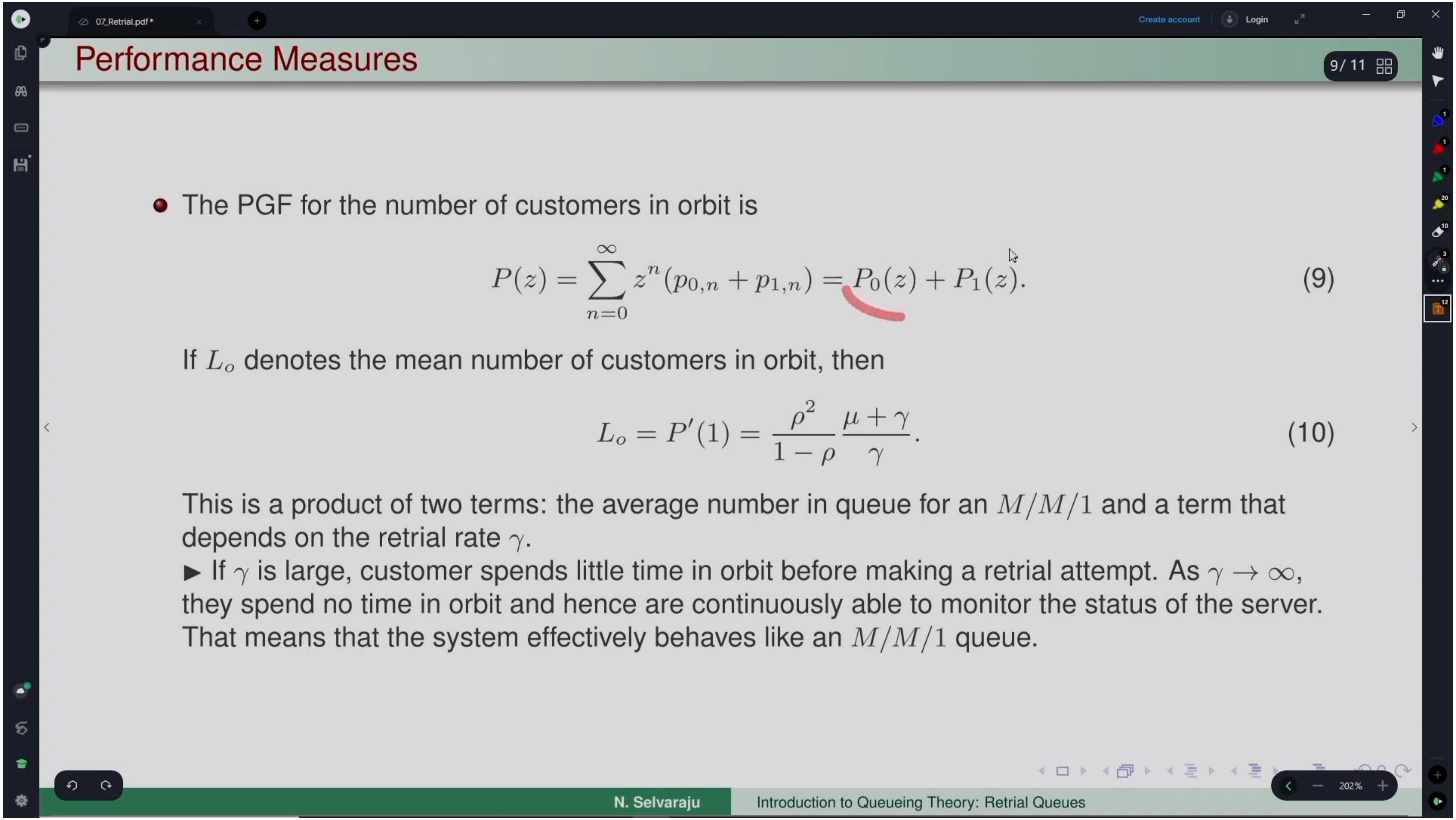


- In a similar manner, one can get $p_{1,n}$ from $P_1(z)$.
- Finally, we obtain the equilibrium probabilities as

$$p_{0,n} = (1-\rho)^{\frac{\lambda}{\gamma}+1} . \frac{\rho^n}{n!\gamma^n} \prod_{i=0}^{n-1} (\lambda + i\gamma), \qquad n \geq 0$$

$$p_{1,n} = (1-\rho)^{\frac{\lambda}{\gamma}+1} . \frac{\rho^{n+1}}{n!\gamma^n} \prod_{i=1}^{n} (\lambda + i\gamma). \qquad n \geq 0.$$

- The fraction of time the server is busy is $\sum_{n=0}^{\infty} p_{1,n} = P_1(1) = \rho$.
  - One can obtain this result by applying Little's law to the server as well.



Performance Measures

- The PGF for the number of customers in orbit is

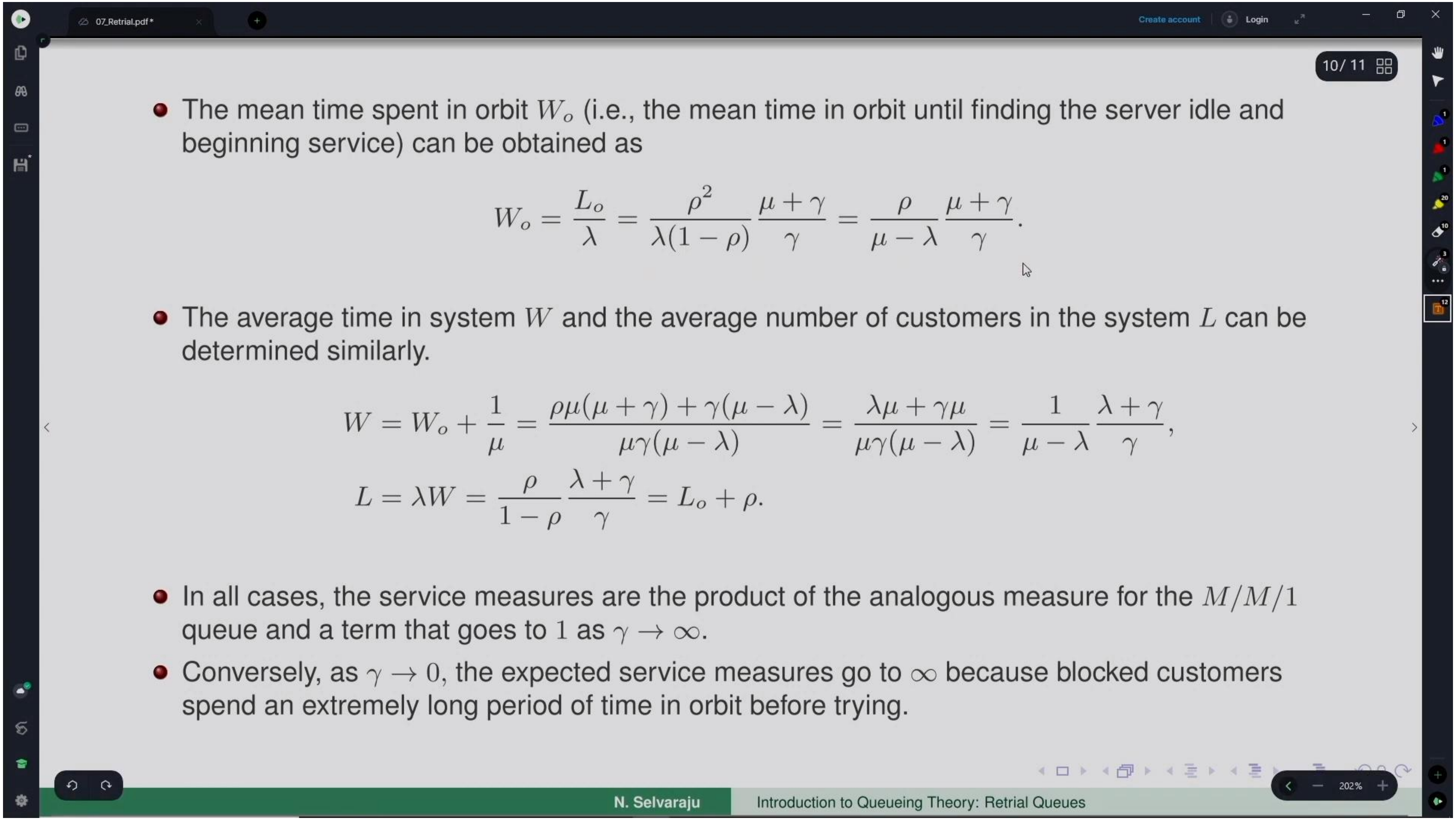
$$P(z) = \sum_{n=0}^{\infty} z^n (p_{0,n} + p_{1,n}) = P_0(z) + P_1(z). \tag{9}$$

If L_o denotes the mean number of customers in orbit, then

$$L_o = P'(1) = \frac{\rho^2}{1-\rho} \frac{\mu + \gamma}{\gamma}. \tag{10}$$

This is a product of two terms: the average number in queue for an $M/M/1$ and a term that depends on the retrial rate γ.

▶ If γ is large, customer spends little time in orbit before making a retrial attempt. As $\gamma \to \infty$, they spend no time in orbit and hence are continuously able to monitor the status of the server. That means that the system effectively behaves like an $M/M/1$ queue.



- The mean time spent in orbit W_o (i.e., the mean time in orbit until finding the server idle and beginning service) can be obtained as

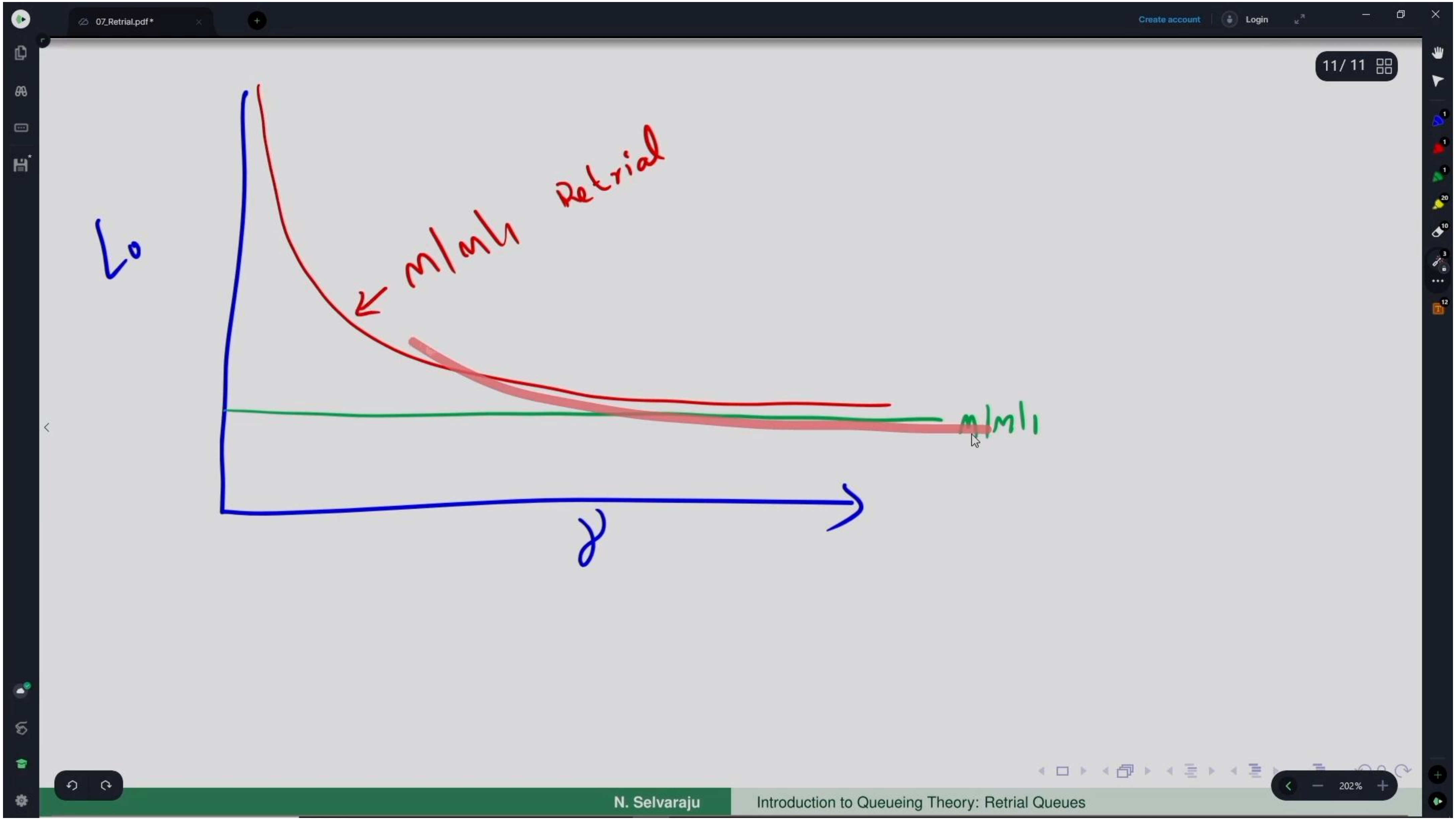
$$W_o = \frac{L_o}{\lambda} = \frac{\rho^2}{\lambda(1-\rho)}\frac{\mu+\gamma}{\gamma} = \frac{\rho}{\mu-\lambda}\frac{\mu+\gamma}{\gamma}.$$

- The average time in system W and the average number of customers in the system L can be determined similarly.

$$W = W_o + \frac{1}{\mu} = \frac{\rho\mu(\mu+\gamma)+\gamma(\mu-\lambda)}{\mu\gamma(\mu-\lambda)} = \frac{\lambda\mu+\gamma\mu}{\mu\gamma(\mu-\lambda)} = \frac{1}{\mu-\lambda}\frac{\lambda+\gamma}{\gamma},$$

$$L = \lambda W = \frac{\rho}{1-\rho}\frac{\lambda+\gamma}{\gamma} = L_o + \rho.$$

- In all cases, the service measures are the product of the analogous measure for the $M/M/1$ queue and a term that goes to 1 as $\gamma \to \infty$.
- Conversely, as $\gamma \to 0$, the expected service measures go to ∞ because blocked customers spend an extremely long period of time in orbit before trying.



N. Selvaraju

Introduction to Queueing Theory: Retrial Queues