Mean Waiting Time Analysis of M/G/1 Queue with Vacations and Gated-Limited Service Discipline

Huanhuan Huang[a], Tong Ye[a], and Tony T. Lee[b]

[a]*Shanghai Jiao Tong University, Shanghai, China*

[b]*Chinese University of Hong Kong (Shenzhen), Shenzhen, China*

August 21, 2017

Outline

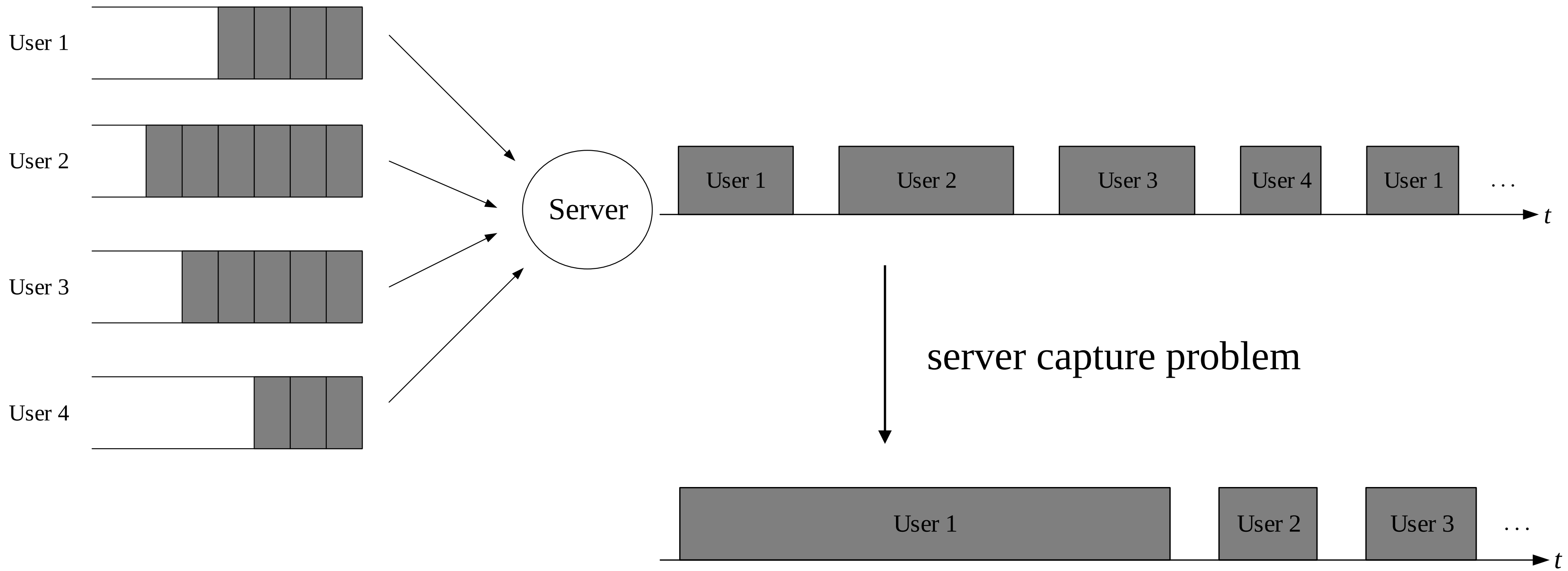
- Problem Description
- System model and Analytical Result
- Conclusion

Outline

- Problem Description
- System model and Analytical Result
- Conclusion

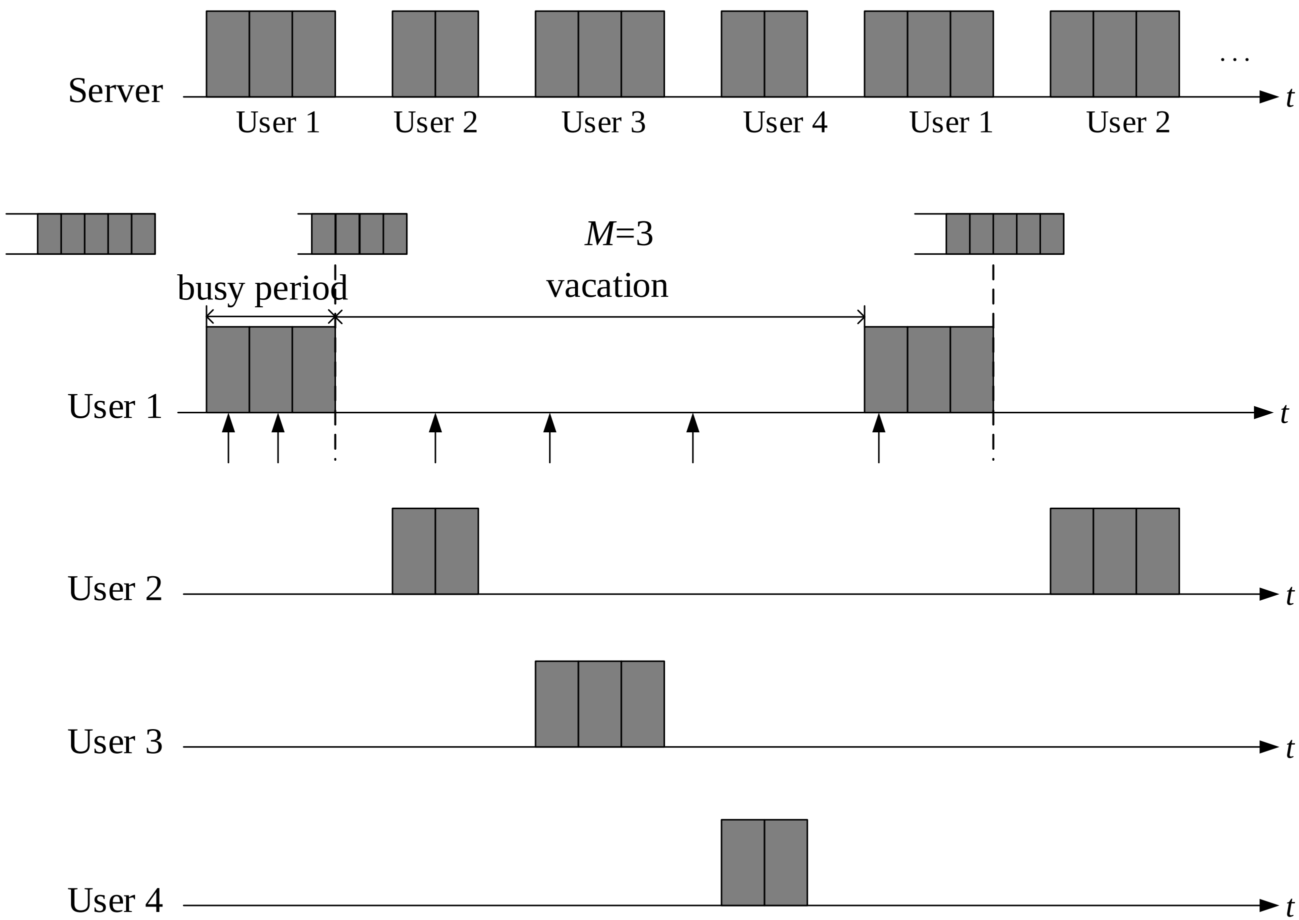
Polling System

- Multiple users share the server in a time-division-multiplexing (TDM) manner
- A proper service discipline is necessary to avoid server capture problem



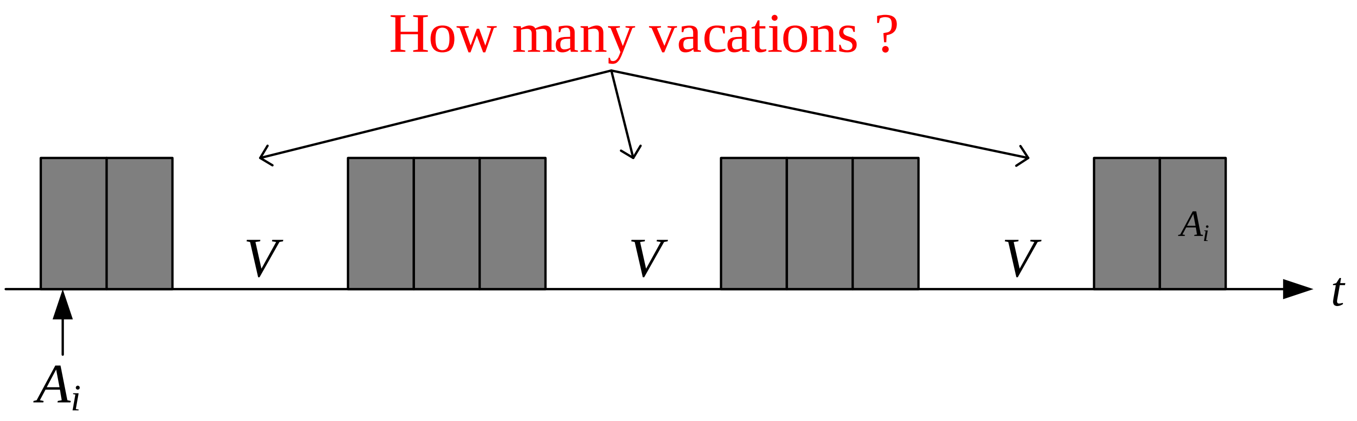
Gated-Limited Service Discipline

- The server serves up to *M* packets that arrived before the end of last busy period



Difficulty in Delay Analysis

- It is hard to analyze the number of vacations that a packet has to experience before it gets served



Previous work & Our goal

- Previous work
  - Embedded Markov chain approach: distributions of queue length, waiting time and busy periods[1-6].
  - Geometric approach: mean waiting time of 1-limited service queuing system[7].
- Our goal
  - To obtain a comprehensive formula of mean waiting time for M/G/1 queue with vacations and gated-limited service discipline where the limit can be arbitrary.

[1] Lee, T. T.: M/G/1/N queue with vacation time and limited service discipline. Performance Evaluation, 9(3), 181–190 (1989).

[2] Van Vuuren, M.: Iterative approximation of k-limited polling systems. Queueing Systems, 55(3), 161–178 (2007).

[3] Al Hanbali, A.: Time-limited and k-limited polling systems: a matrix analytic solution. In: Proceedings of the 3rd International Conference on Performance Evaluation Methodologies and Tools, p. 17. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Athens, Greece (2008).

[4] Sikha, M. B.: A two-queue finite-buffer polling model with limited service and state-dependent service times. In: IEEE International Conference on Electronics, Computing and Communication Technologies (IEEE CONECCT), pp. 1–6. IEEE, Bangalore, India (2014).

[5] Takagi, H.: Queueing Analysis: A Foundation of Performance Analysis, Volume 1: Vacation and Priority Systems. 1st edn. Elsevier, Amsterdam (1991).

[6] Alfa, A. S.: Vacation models in discrete time. Queueing Systems, 44(1), 5–30 (2003).

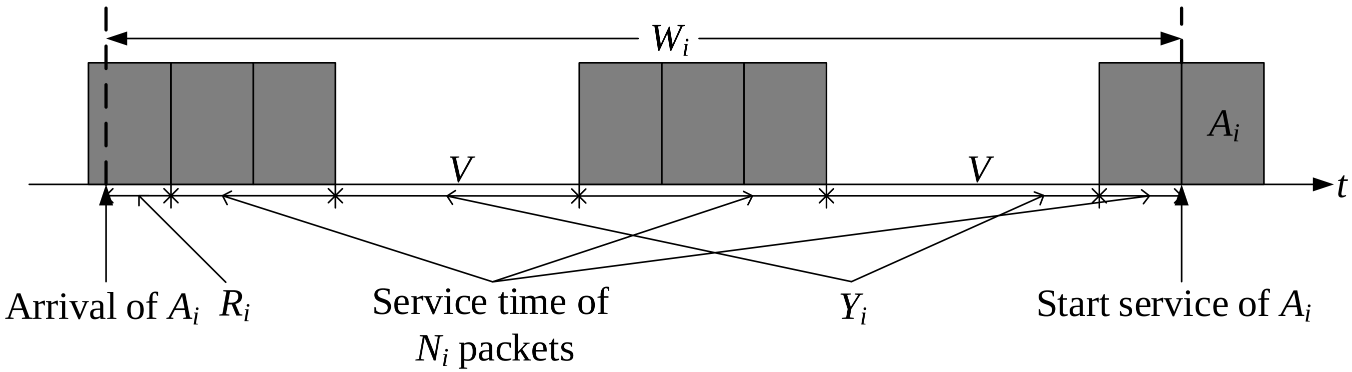
[7] Bertsekas, D. P., Gallager, R. G., Humblet, P.: Data networks. 2nd edn. Prentice-Hall In-ternational, New Jersey (1992).

Outline

- Problem Description
- System model and Analytical Result
- Conclusion

Mean Waiting Time Analysis

- Waiting time W_i of customer A_i consists of:
  - Residual service or vacation time R_i
  - Service time for N_i packets waiting before A_i
  - Whole vacation periods Y_i experienced by A_i before it gets service



$$\bar{W} = E[W_i]$$

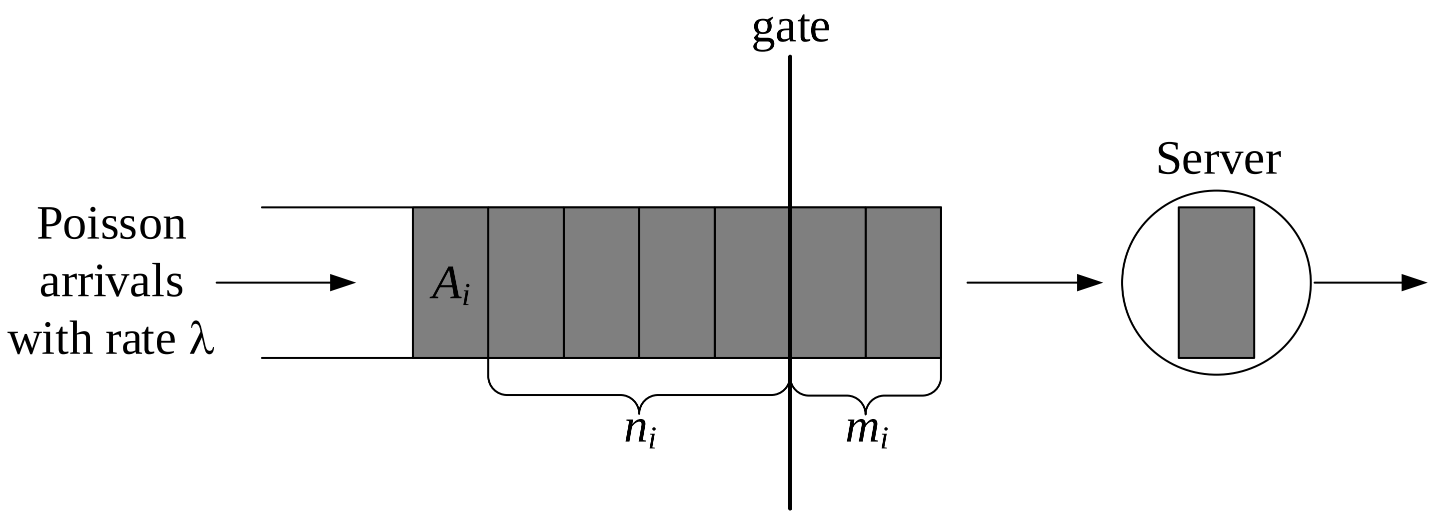
$$= E[R_i] + E\left[\sum_{j=i-N_i}^{i-1} X_j\right] + E[Y_i]$$

$$= \bar{R} + N_Q\bar{X} + \bar{Y}$$

[7] D. P. Bertsekas, R. G. Gallager and P. Humblet, “Data networks”. Prentice-Hall International, 1992.

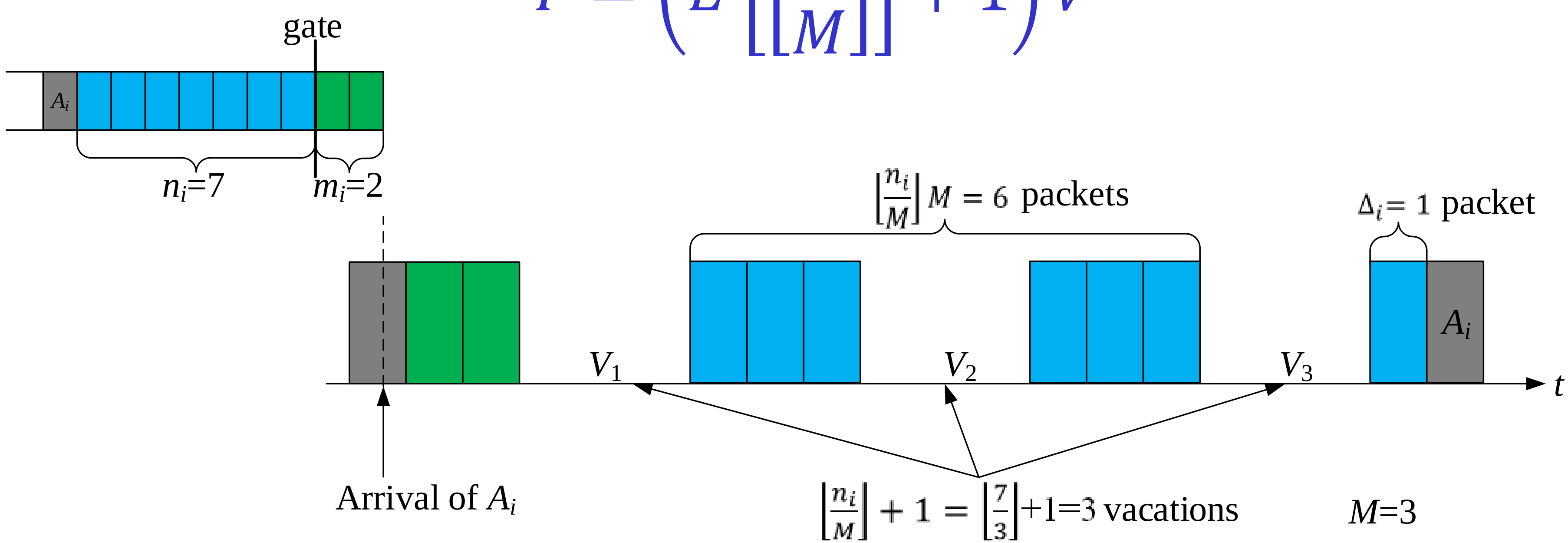
Two-Queue Buffer Model

- Buffer inside the gate: m_i $(\leq M)$
- Buffer outside the gate: n_i



Components of $\bar{Y}$

$$\bar{Y} = \left(E\left[\left\lfloor\frac{n_i}{M}\right\rfloor\right] + 1\right)\bar{V}$$



$$n_i = \left\lfloor\frac{n_i}{M}\right\rfloor M + \Delta_i$$

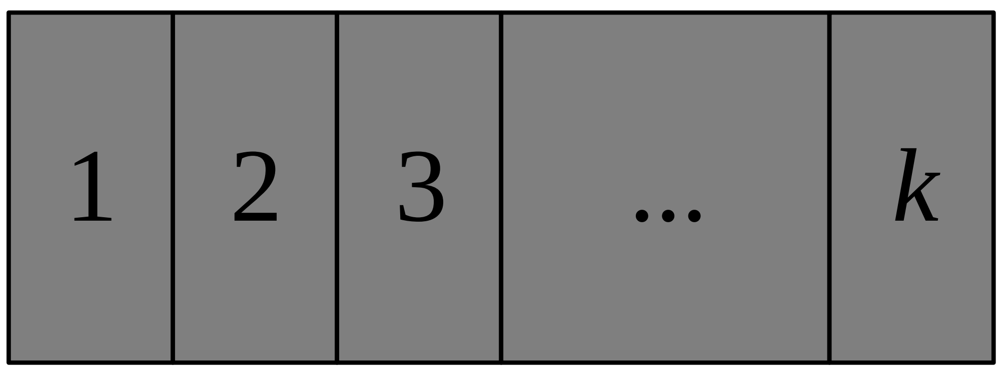
Number of packets served immediately before A_i in the same busy period

$$E\left[\left\lfloor\frac{n_i}{M}\right\rfloor\right] = \frac{E[n_i] - E[\Delta_i]}{M}$$

Conditional Expectation $E[\Delta_i|k]$

- F_k: packet A_i is served in a busy period of length k
- $E[\Delta_i|k]$: mean number of packets served immediately before A_i given that F_k happens

$$E[\Delta_i|k] = \frac{0+1+\cdots+k-1}{k} = \frac{k-1}{2}$$



$\Delta_1=0$ $\Delta_2=1$ $\Delta_3=2$ $\cdots$ $\Delta_k=k-1$

Probability that F_k Happens

- $B_k (k = 0,1, \dots , M)$: a busy period of length k
- $K = k$: number of packets served in a busy period
- $b_k = \Pr\{\text{a busy period is a } B_k\}$

$$Pr\{F_k \text{ happens}\}$$
$$= \lim_{T\to\infty} \frac{\text{number of packets served in } B_k\text{s within } [0, T]}{\text{number of packets served within } [0, T]}$$
$$= \frac{k b_k}{\sum_{k=1}^{M} k b_k}$$
$$= \frac{k b_k}{\overline{K}}$$

Expectation $E[\Delta_i]$

$$E[\Delta_i] = E\big[E[\Delta_i|k]\big]$$

$$= \sum_{k=1}^{M} E[\Delta_i|k]\, Pr\{F_k\}$$

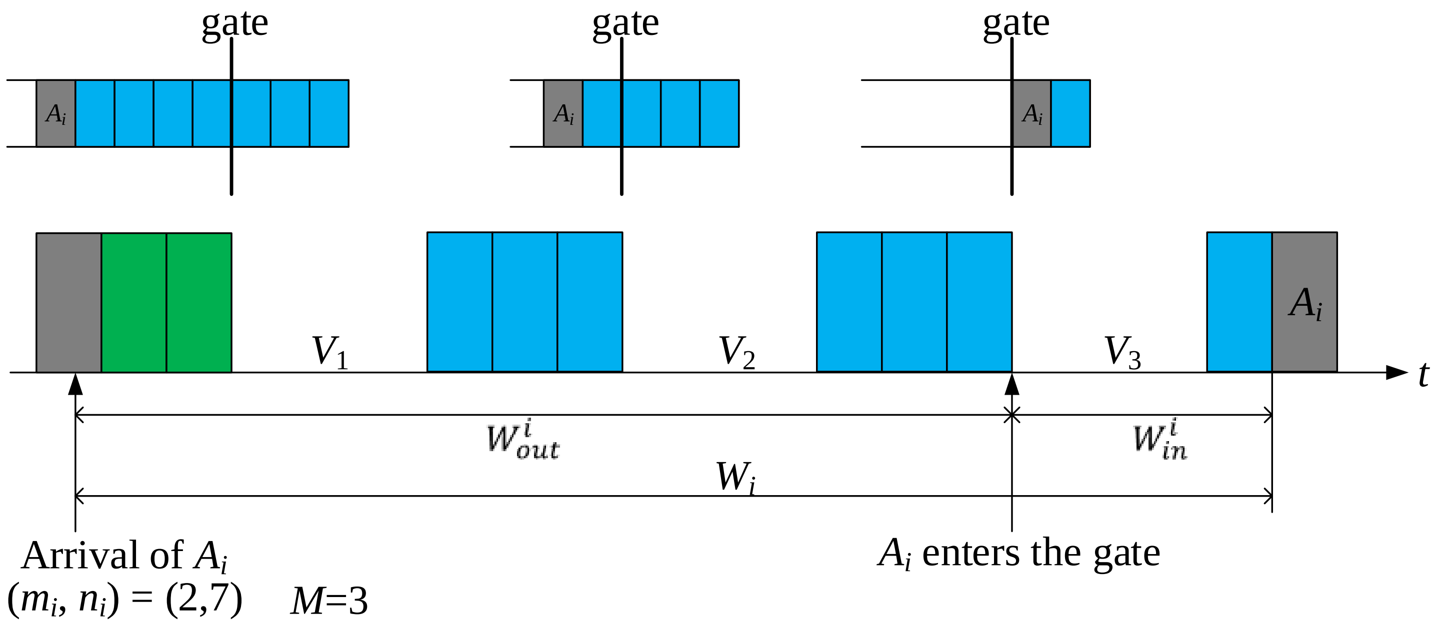
$$= \sum_{k=1}^{M} \frac{k-1}{2} \cdot \frac{k b_k}{\overline{K}}$$

$$= \frac{\overline{K^2} - \overline{K}}{2\overline{K}}$$

Mean Queue Length Outside the Gate $E[n_i]$

- W_{out}^i: waiting time of A_i outside the gate
- W_{in}^i: waiting time of A_i inside the gate

$$E[n_i] = \lambda E\left[W_{out}^i\right] = \lambda\left(\bar{W} - E\left[W_{in}^i\right]\right) = \lambda[\bar{W} - (\bar{V} + E[\Delta_i]\bar{X})]$$



Mean Duration of Whole Vacation Periods $\bar{Y}$

$$\bar{Y} = \left(E\left[\left\lfloor\frac{n_i}{M}\right\rfloor\right] + 1\right)\bar{V}$$

$$= \left(\frac{E[n_i] - E[\Delta_i]}{M} + 1\right)\bar{V}$$

$$= \left[\frac{\lambda\bar{W} - \lambda\bar{V}}{M} - \frac{(1+\rho)\left(\overline{K^2} - \bar{K}\right)}{2M\bar{K}} + 1\right]\bar{V}$$

Mean Waiting Time Formula

- **Theorem 1:** The mean waiting time of M/G/1 queue with vacations and gated-limited service discipline is given by

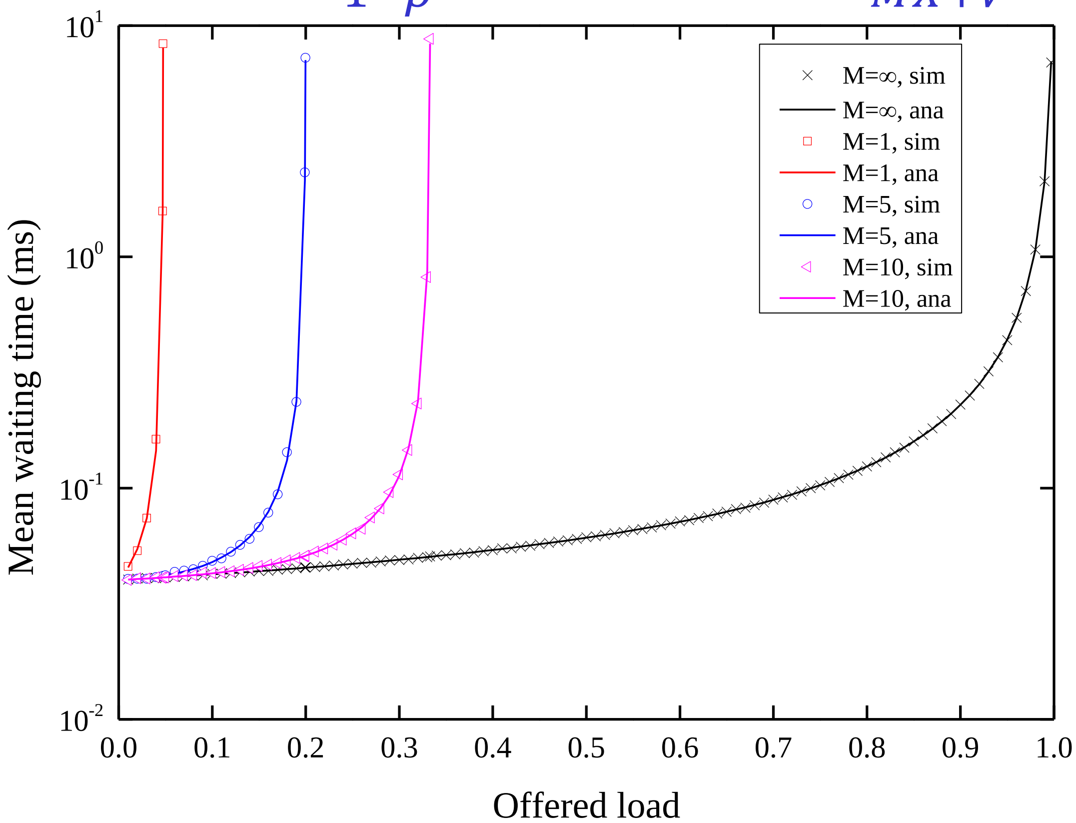
$$\bar{W} = \frac{\frac{\lambda \overline{X^2}}{2} + \frac{(1-\rho)\overline{V^2}}{2\bar{V}} + \left[1 - \frac{(1+\rho)\left(\overline{K^2} - \bar{K}\right)}{2M\bar{K}} - \frac{\lambda \bar{V}}{M}\right]\bar{V}}{1 - \rho - \frac{\lambda \bar{V}}{M}}.$$

- $\overline{K^2}$: second moment of the number of packets served in a busy period
- $\overline{K^2}$ can be derived by using an embedded Markov chain, where vacation start points are considered as embedded points

[5] Takagi, H.: Queueing Analysis: A Foundation of Performance Analysis, Volume 1: Vacation and Priority Systems. 1st edn. Elsevier, Amsterdam (1991).

Simulation vs. Analytical Results

- Packet service time and vacation time are assumed exponentially distributed with $\bar{X} = 1\mu s$ and $\bar{V} = 20\mu s$ respectively.
- Stable condition: $\frac{\lambda \bar{V}}{1-\rho} < M \longrightarrow \rho < \frac{M\bar{X}}{M\bar{X}+\bar{V}}$



Outline

- Problem Description
- System model and Analytical Result
- Conclusion

Conclusion

- We derive a comprehensive formula of mean waiting time for M/G/1 queue with vacations and gated-limited service discipline
- We find that the mean delay is related to the first and second moment of service time, vacation time and the number of packets served in a busy period

Thank You!