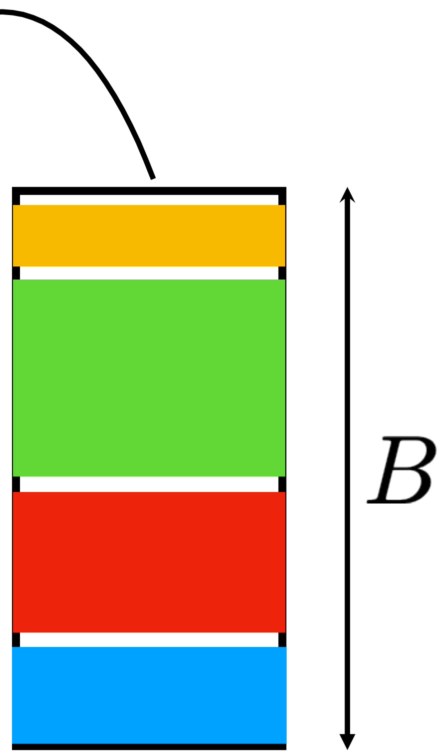
Speed Scaling under QoS Constraints with Finite Buffer

Parikshit Hegde, Akshit Kumar, and Rahul Vaze

Motivation:Example

packets arrivals A_t **i.i.d**



Server

packets/sec

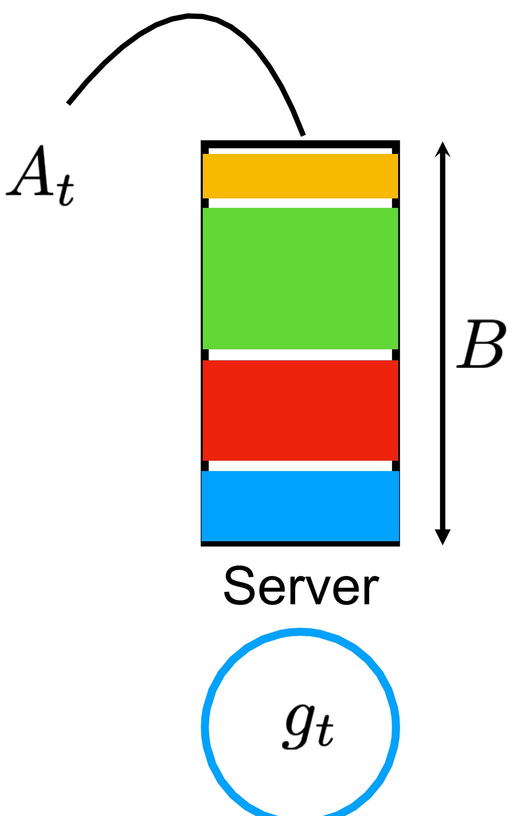
cost $f(g_t)$

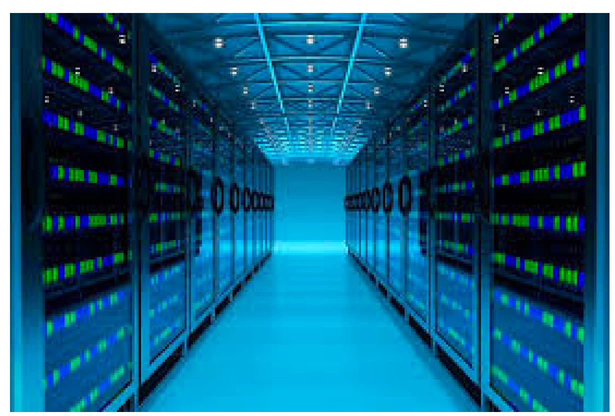
Constraint $P_{\text{drop}} \leq \alpha$

Obj: min $\frac{1}{n}\sum_{t=1}^{n} f(g_t)$

Prior Work

- Problem formulated *Srikant &Perkins*, '99
- Computing Optimal Policies is challenging
- Simplified the problem
  - We propose near optimal policies with provable guarantees





$$\frac{1}{n}\sum_{t=1}^{n} f(g_t) + \lambda P_{\mathrm{drop}}$$

cost $f(g_t)$

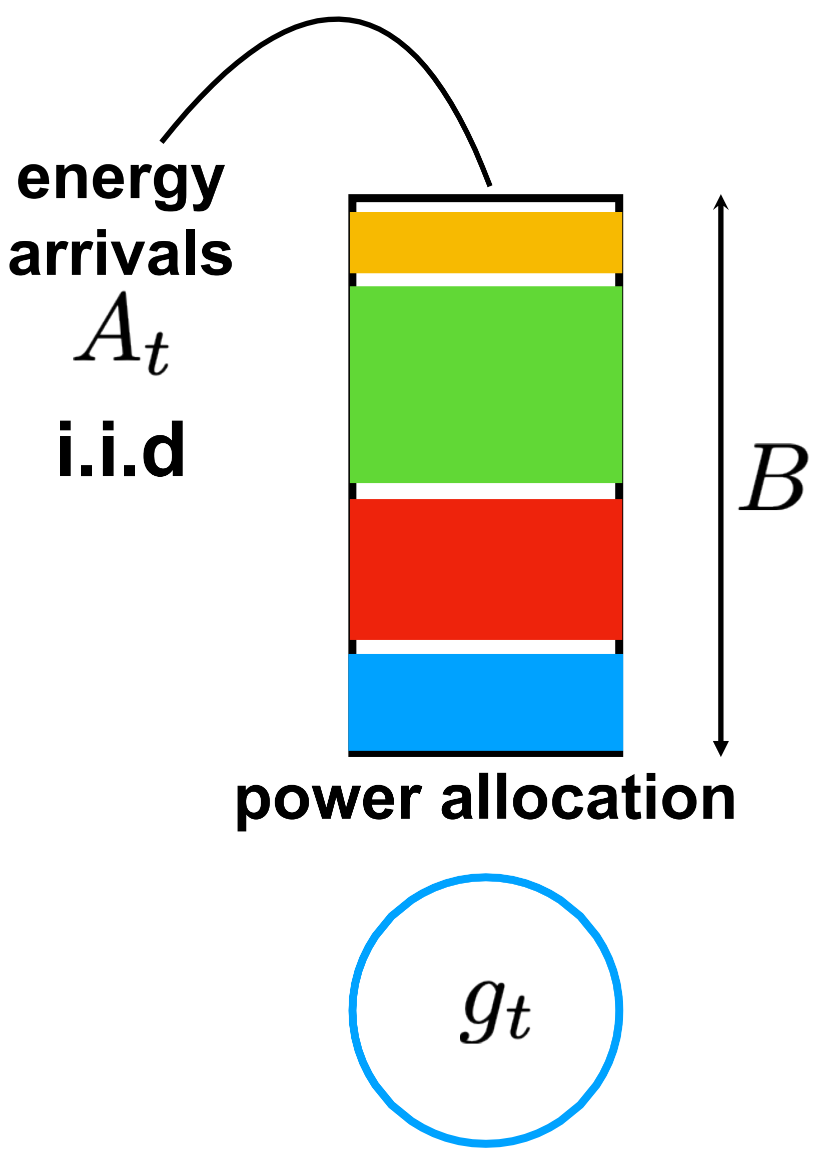
Add as penalty

Constraint $P_{\mathrm{drop}} \leq \alpha$

Obj: min $\frac{1}{n}\sum_{t=1}^{n} f(g_t)$

More Recent Work - EH

- [*Shaviv & Ayfer*, 2016] Competitive ratio of 2
- If B is large, [*Srivastava & Koksal*, 2013] provide near optimal policies
- In addition they show that battery overflow or battery discharge $\Theta\left(B^{-\beta}\right)$

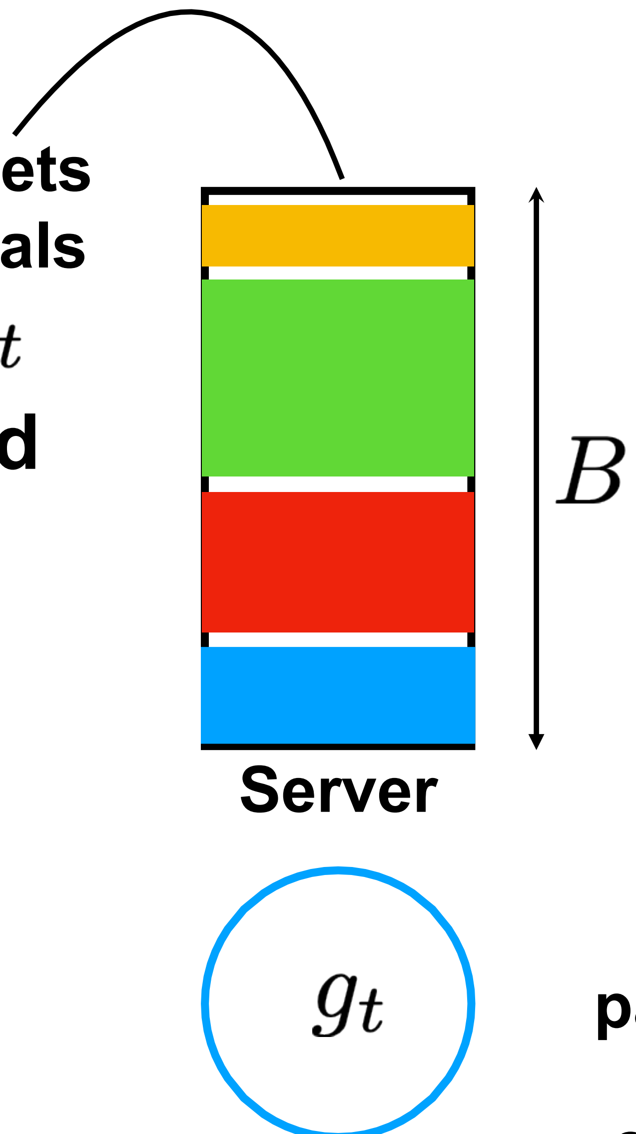


Utility Max $$\frac{1}{n}\sum_{t=1}^{n} U(g_t)$$

concave

Recap - Problem

packets arrivals
A_t
i.i.d



packets/sec

cost $f(g_t)$

Constraint $P_{\text{drop}} \leq \alpha$

Obj: min $\frac{1}{n}\sum_{t=1}^{n} f(g_t)$

General Lower Bound

$$P_{\text{drop}} \leq \alpha$$

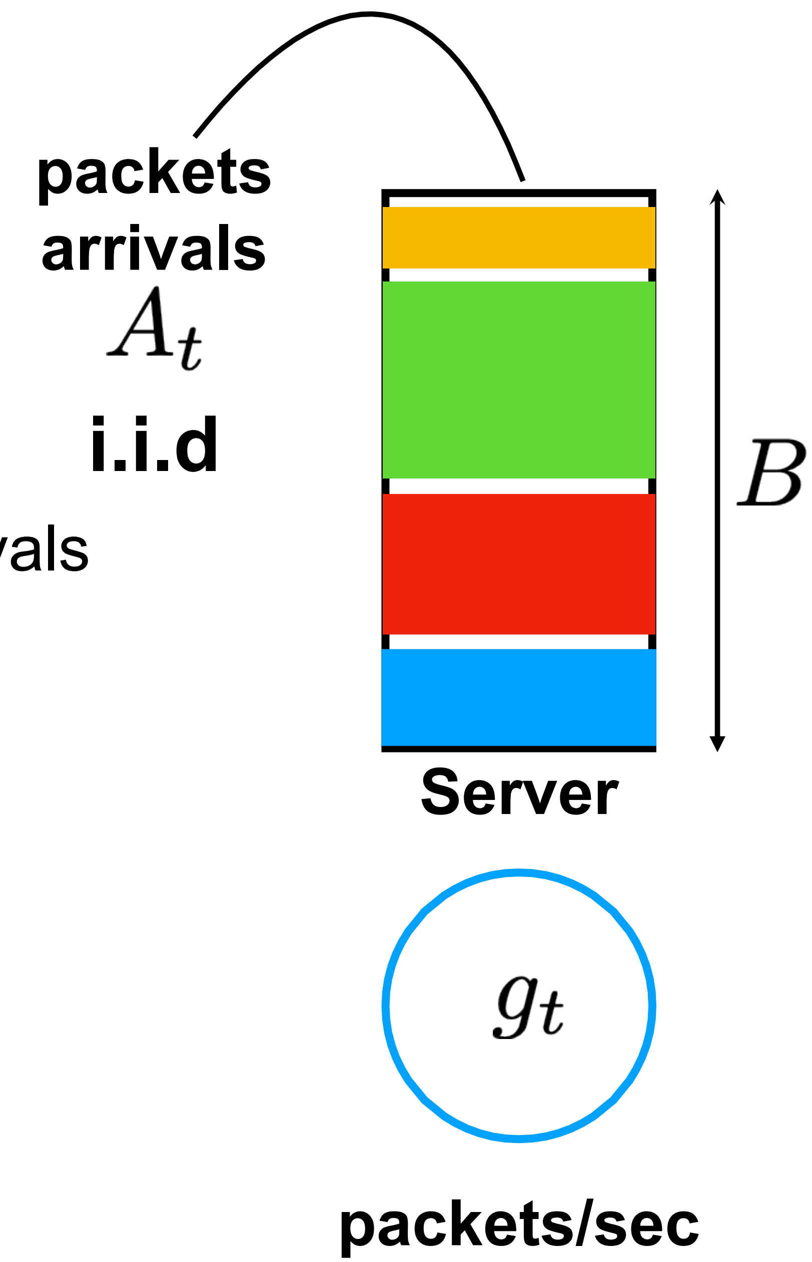
Service at least $(1-\alpha)$ fraction of total packet arrivals

$$\mathbf{E}\left[\frac{\sum_{t=1}^{n} g_t}{n}\right] \geq (1-\alpha)\mathbf{E}\left[\frac{\sum_{t=1}^{n} A_t}{n}\right]$$

$$\lim_{n\to\infty} \mathbf{E}\left[\frac{\sum_{t=1}^{n} g_t}{n}\right] \geq (1-\alpha)\mu$$

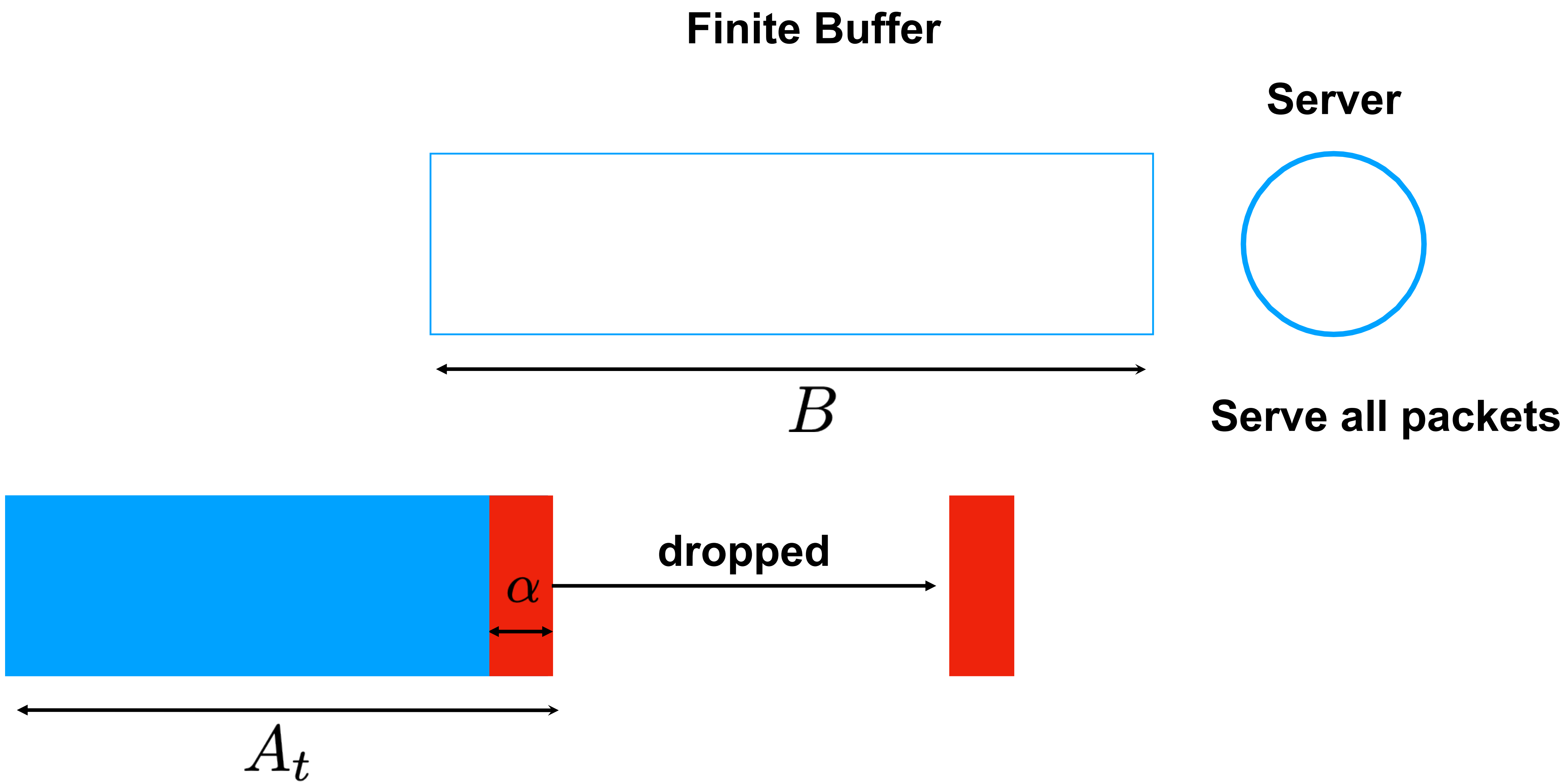
Invoking Jensen's

$$\text{Cost} \geq f((1-\alpha)\mu)$$



Greedy Policy

- Forcefully drop α fraction of arrivals
- Immediately serve the $1 - \alpha$ fraction of arrivals



Greedy Policy:Performance

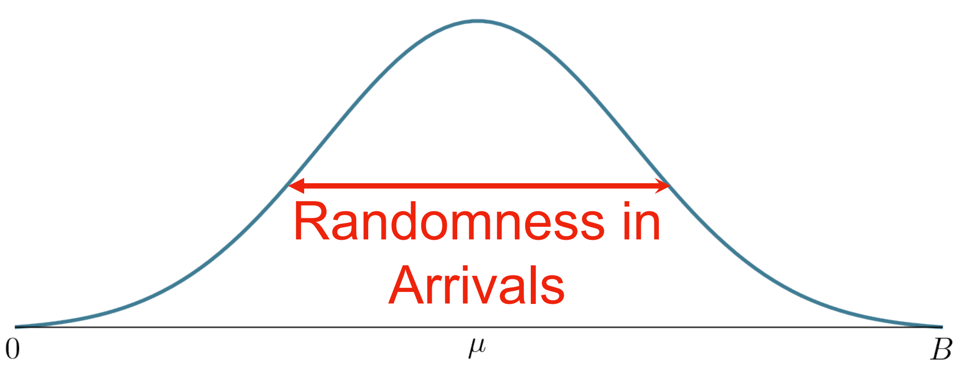
$f(x) = x^2$

- Immediately serve the $1-\alpha$ fraction of arrivals

$$\text{Cost} \leq \lim_{n\to\infty} \mathbf{E}\left[\frac{1}{n}\sum_{t=1}^{n}(1-\alpha)^2 A_t^2\right]$$

$$\leq (1-\alpha)^2\left(\mu^2 + \text{var}\left(A_t\right)\right)$$

Example Arrival Distribution



Total Cost : $(1-\alpha)^2\left(\boxed{\mu^2} + \boxed{var(A_t)}\right)$

Term from Lower Bound

Second Order Moment

sufficient in practice

Greedy Policy:Performance

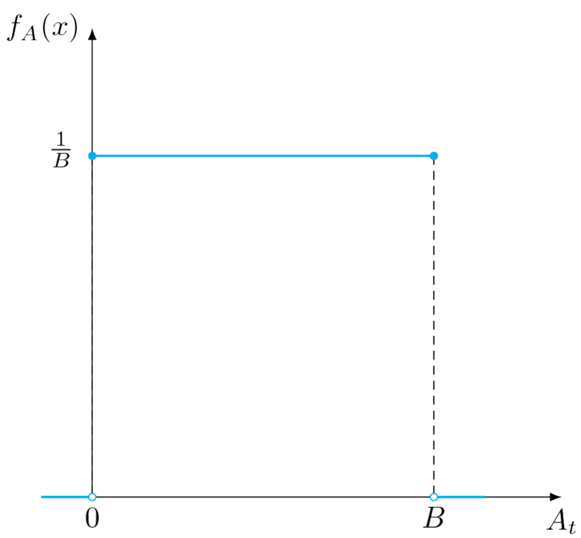
Competitive Ratio(CR) = $1 + \mathrm{var}(A_t)/\mu^2$

Lower Bound
$(1-\alpha)^2\mu^2$

Greedy Policy
$(1-\alpha)^2(\mu^2 + var(A_t))$

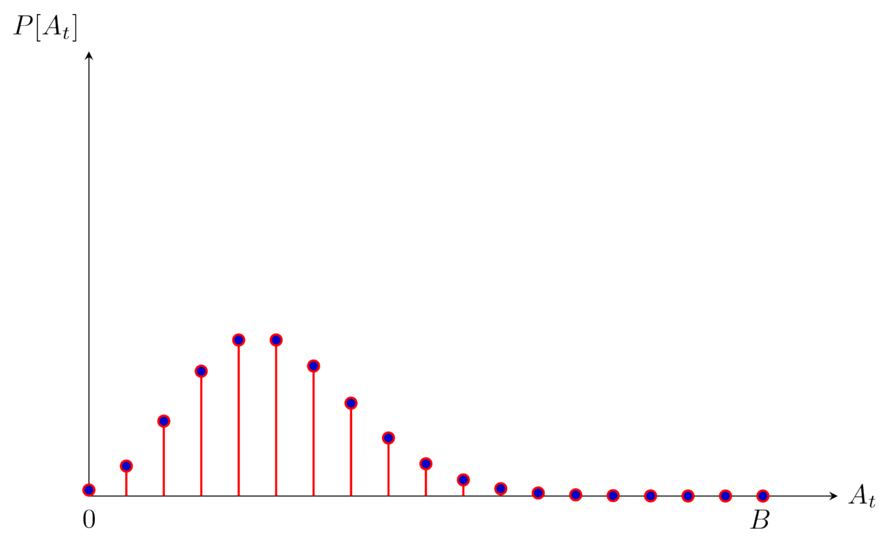
Greedy Policy:Performance

Competitive Ratio(CR) = $1 + \mathrm{var}(A_t)/\mu^2$



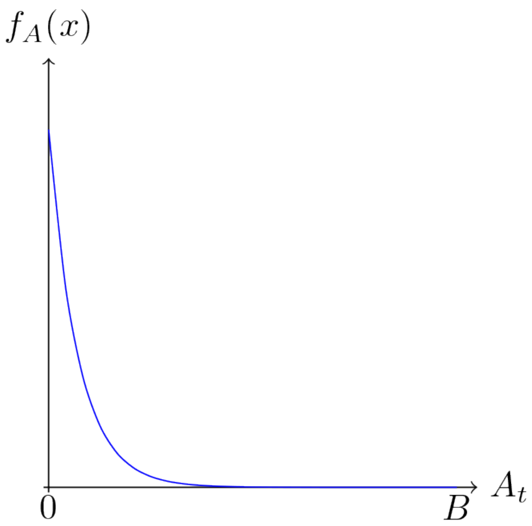
$A_t \sim \mathrm{Unif}[0, B]$

$\mathrm{CR} \leq \frac{4}{3}$



$A_t \sim \mathrm{Poisson}(\nu), \nu \geq 1$

$\mathrm{CR} \leq 3$



$A_t \sim \mathrm{Exp}(\mu)$

$\mathrm{CR} \leq 2$

Greedy Policy:Performance

Bernoulli

$$A_t = \begin{cases} m, & \text{w.p. } \frac{\mu}{m}, \\ 0, & \text{w.p. } 1 - \frac{\mu}{m}, \end{cases}$$

$$0 < m \le B$$

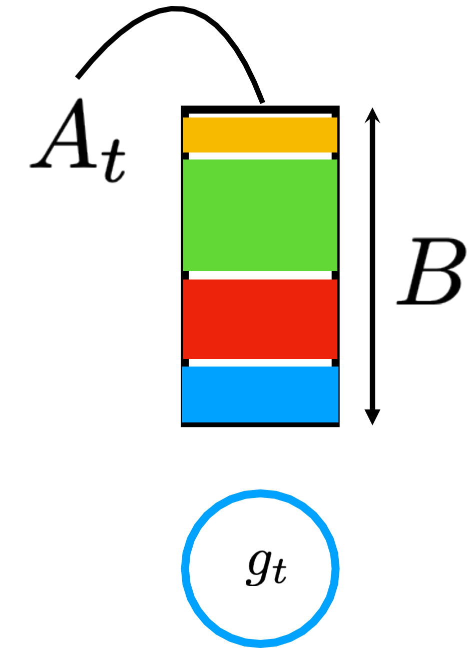
$$\mathrm{CR}_{\mathrm{greedy}} \le 1 + \frac{\mathrm{var}[X]}{\mu^2} = \frac{m}{\mu}$$

$$\mu \to 0$$

Special Case Analysis

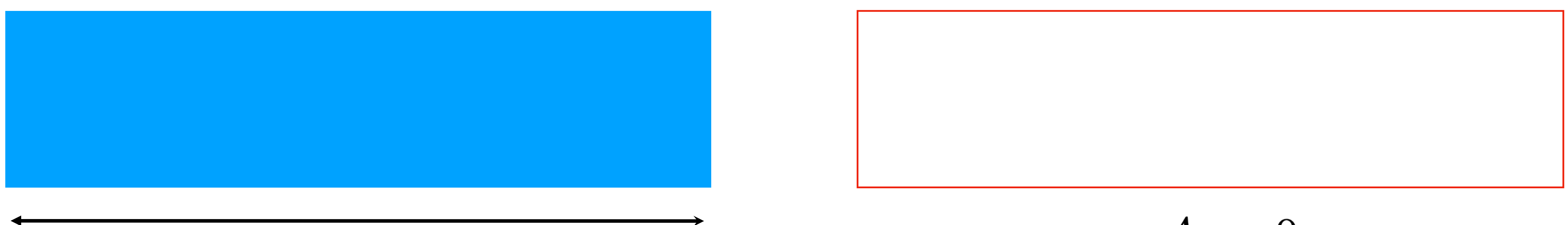
$$\mu \to 0$$

Small μ regime



- Focus on *Extreme Bernoulli Distribution*

With probability p | **With probability** $1-p$



$A_t = B$ | $A_t = 0$

- Why this distribution?

- Conjecture that this is the worst case

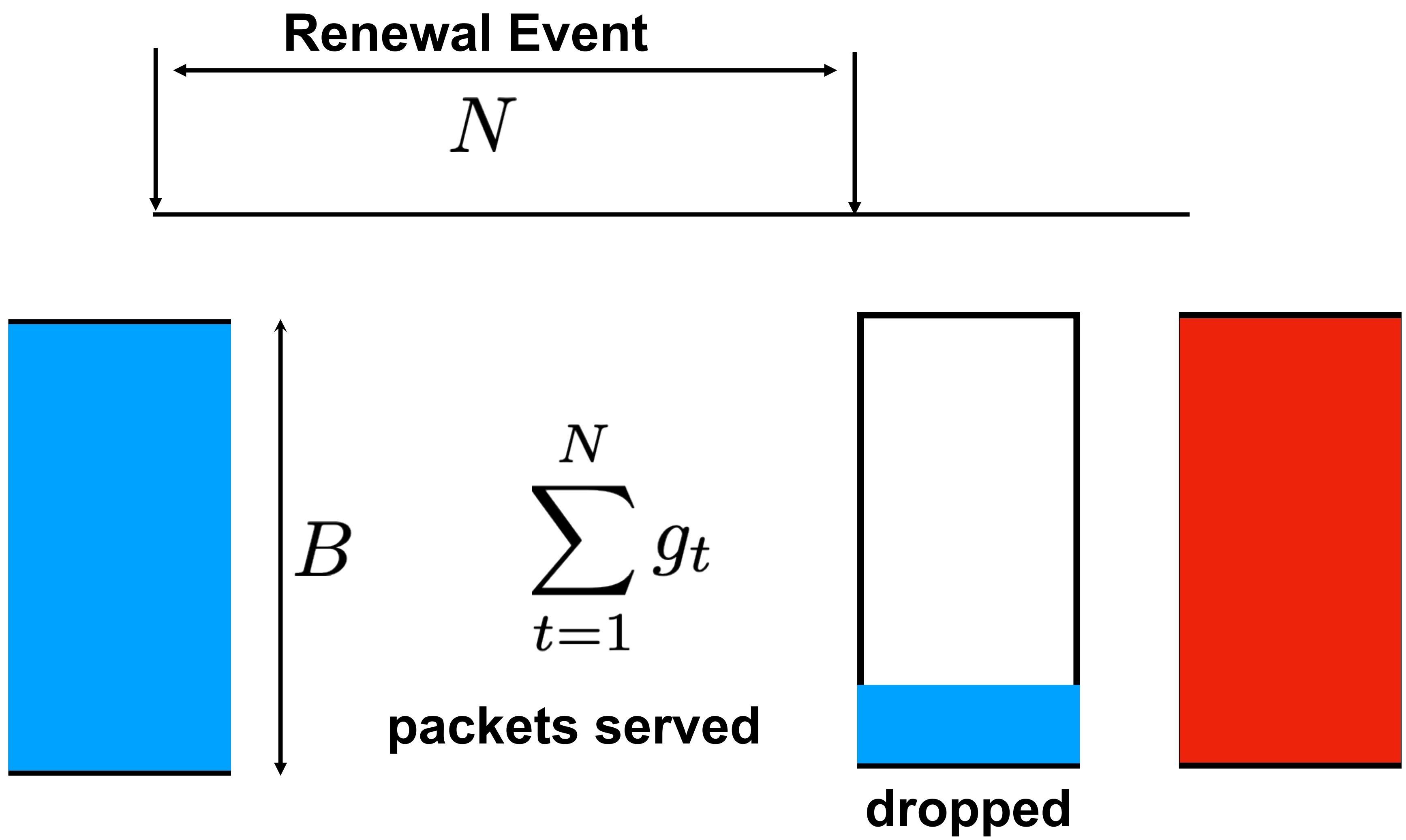
- For similar problem in Energy Harvesting, Shaviv and Ayfer, 2016 showed this is worst case distribution

Extreme Bernoulli Dist

$$P_{\text{drop}} = \frac{\mathbf{E}\{B - \sum_{t=1}^{N} g_t\}}{B}$$

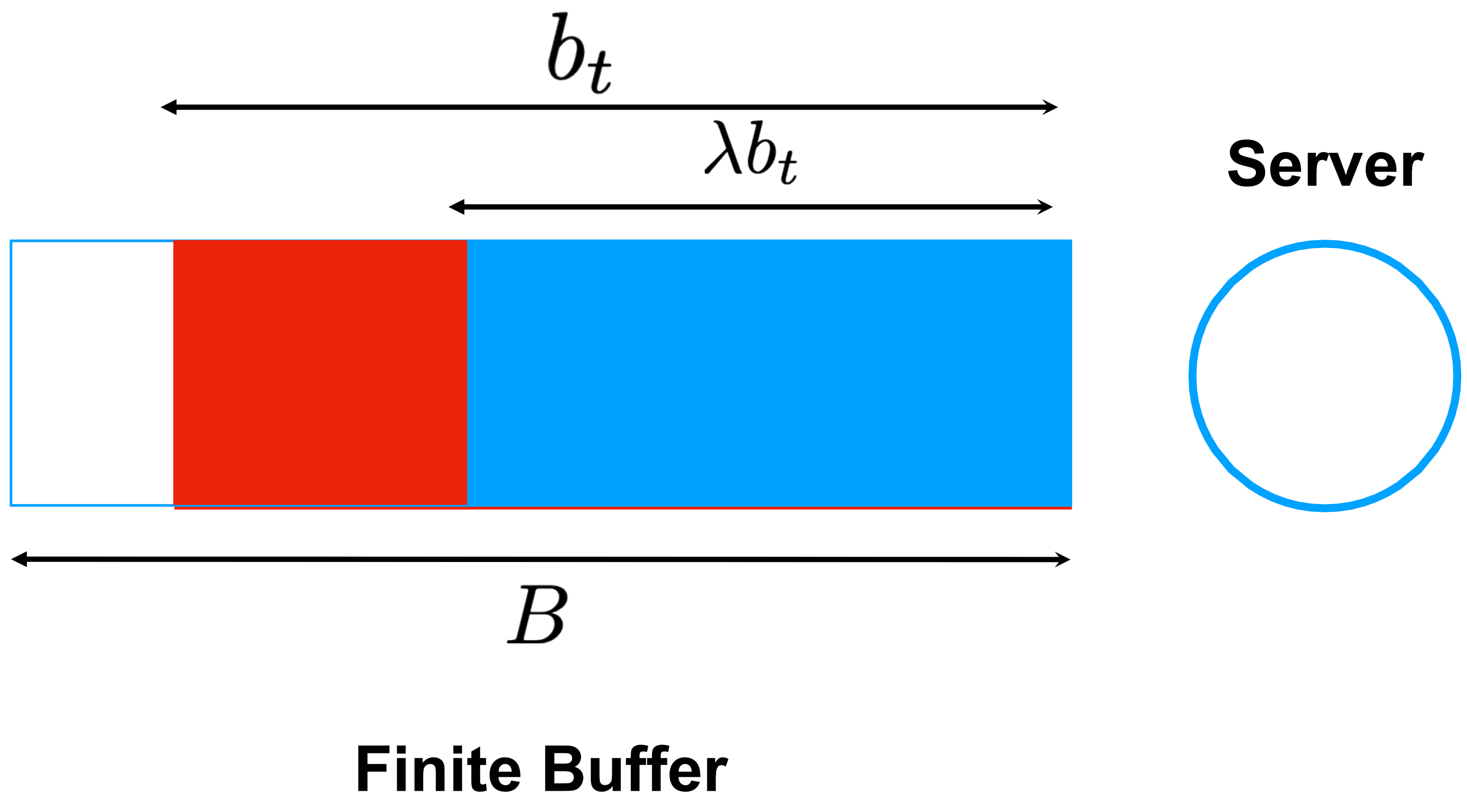
$$P_{\text{drop}} \le \alpha \qquad g_1 \ge B\left(1 - \frac{\alpha}{p}\right)$$

$$\text{Cost} = \frac{\mathbf{E}\left[\sum_{t=1}^{N} g_t^2\right]}{\mathbf{E}[N]} \ge \frac{g_1^2}{\mathbf{E}[N]} \ge \frac{B^2 p^2}{\alpha + p}$$



λ -fraction Policy

- Serve λ fraction of the packets in the buffer
- λ calculated as a function of p to satisfy P_{drop} of α



Performance

$$\text{For small } \alpha << 1, p << 1$$

$$\mathcal{J}^* = \Omega\left(\frac{\mu^2}{p+\alpha}\right)$$

Lower Bound

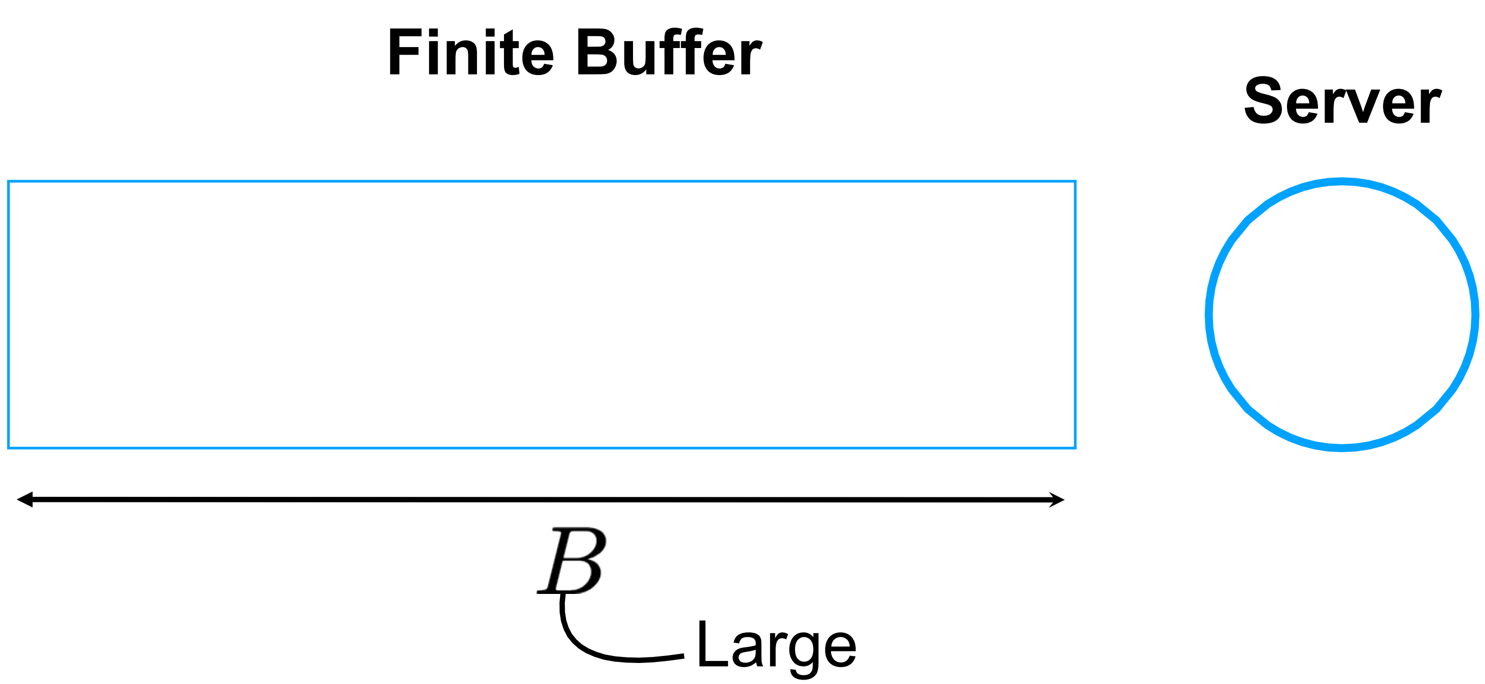
$$\mathcal{J}_\lambda \approx \frac{\mu^2}{2\alpha + p}$$

λ Fraction Policy

Large Buffer Case

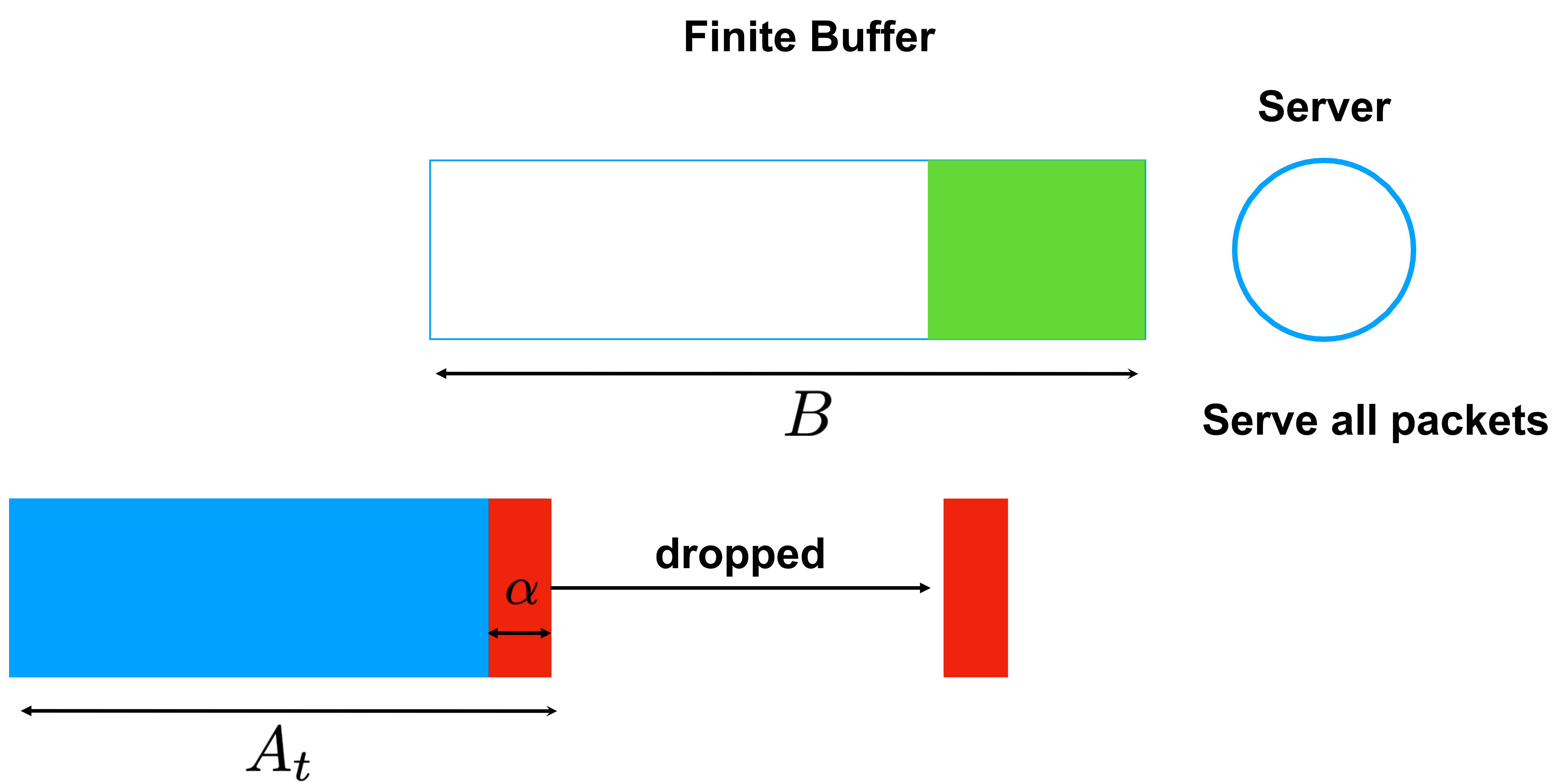
Large Buffer Case

- Consider the case where buffer size is large
- Idea: Large buffer will reduce the "effect" of randomness in arrivals
- General Lower Bound that assumes μ arrivals every time likely to be achievable



Admission Policy

- Forcefully drop α fraction of arrivals

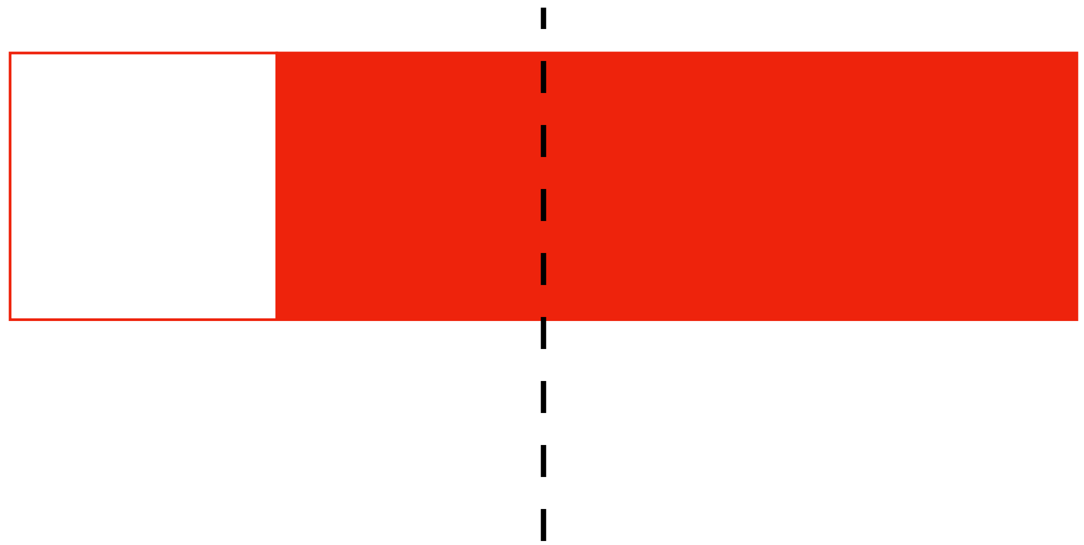


Scheduling Policy

Buffer/Queue state b_t

$$b_t > B/2$$

Finite Buffer



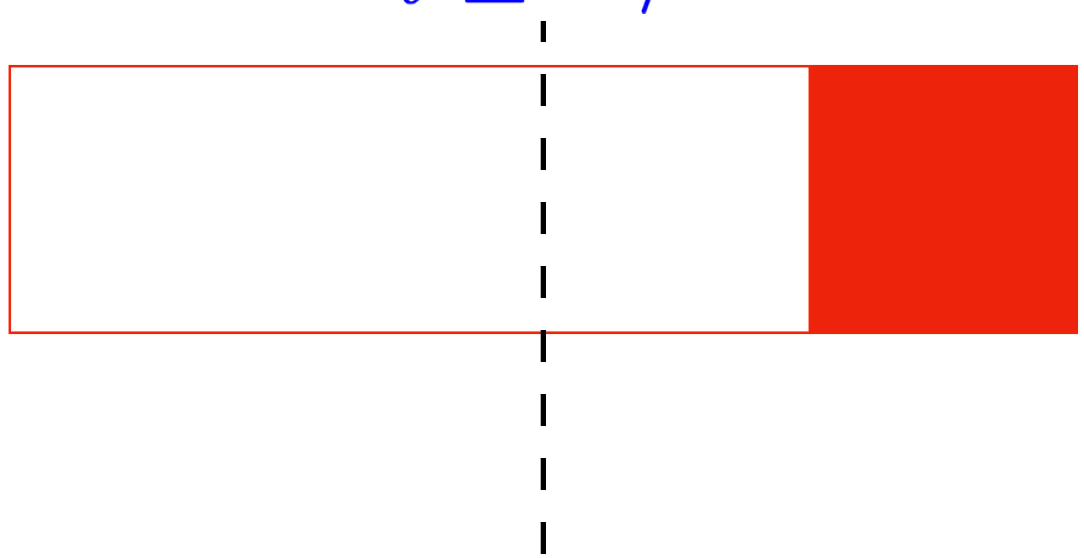
$$g_t = \mu(1-\alpha) + \delta$$

$B/2$

δ **maintain Packet Drop constraint**

$$\delta \quad \sim \beta\left(\frac{\log B}{B}\right)$$

$$b_t \leq B/2$$



$$g_t = \mu(1-\alpha) - \delta$$

Results

$$\text{Cost} \leq \text{Opt} + \Theta(B^{-\beta}) + \Theta\left(\beta^2\left(\frac{\log B}{B}\right)^2\right)$$

General Lower Bound

β **tradeoffs the two terms**

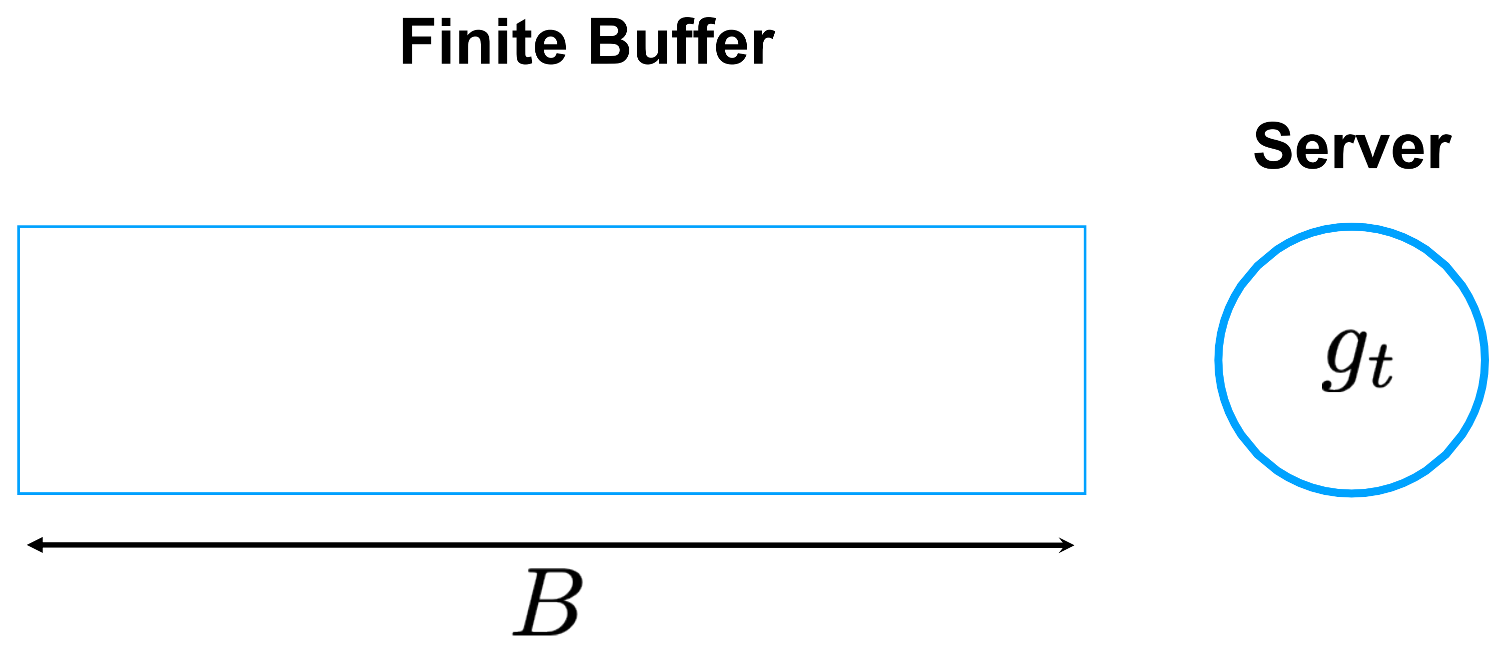
$$P_{\text{drop}} \leq \alpha + \Theta(B^{-\beta})$$

Small Violation to the Packet Drop Constraint

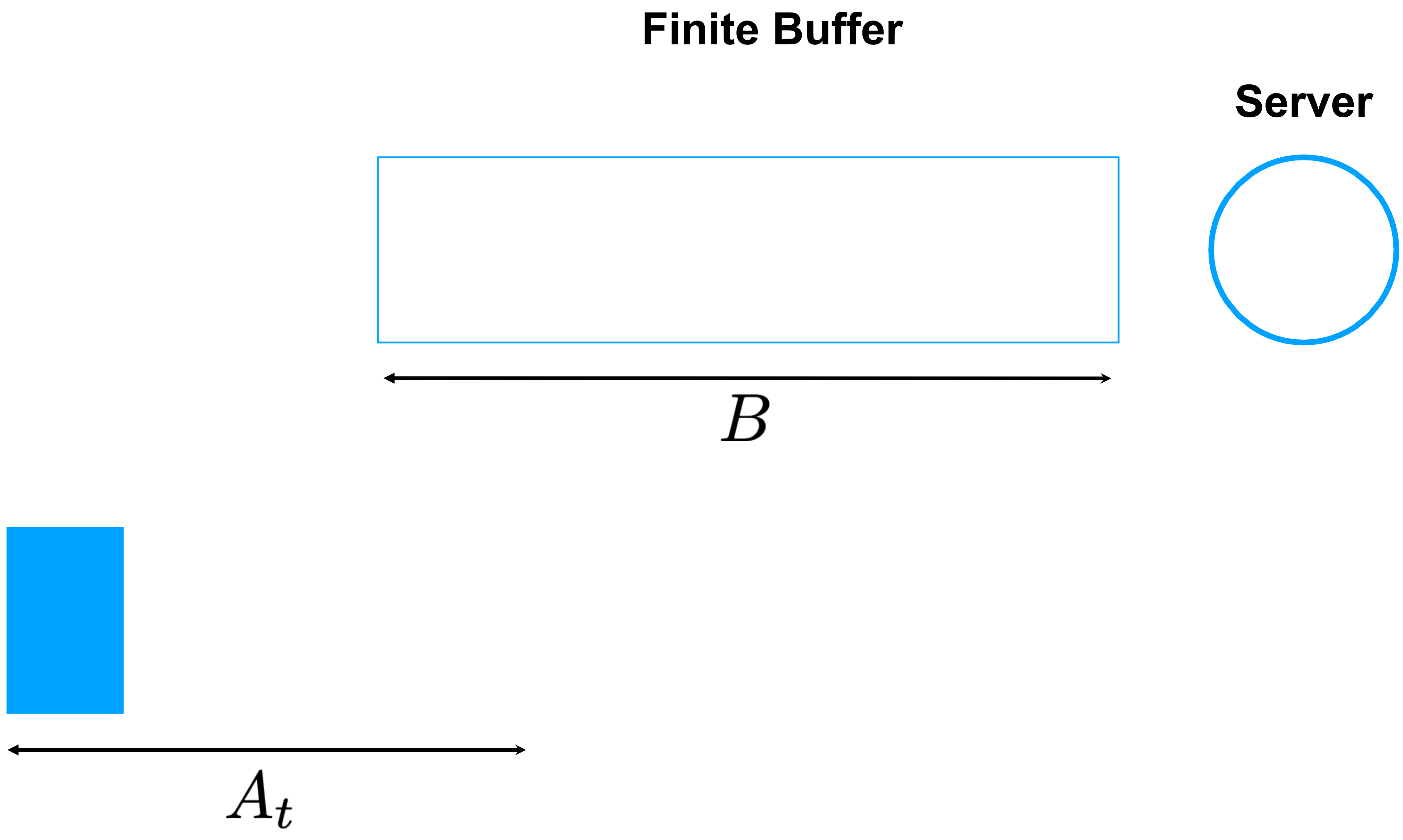
Summary

- Problem of Speed Scaling a server
    - QoS constraint imposed by Packet Drop
    - Minimise total cost
- Greedy Policy is near optimal when ***var to mean ratio*** is small
- For low arrival rate
    - Focus on *extreme Bernoulli distribution*
    - λ-fraction policy is near optimal
- For more general distributions
    - Near Optimal policy when buffer size is large

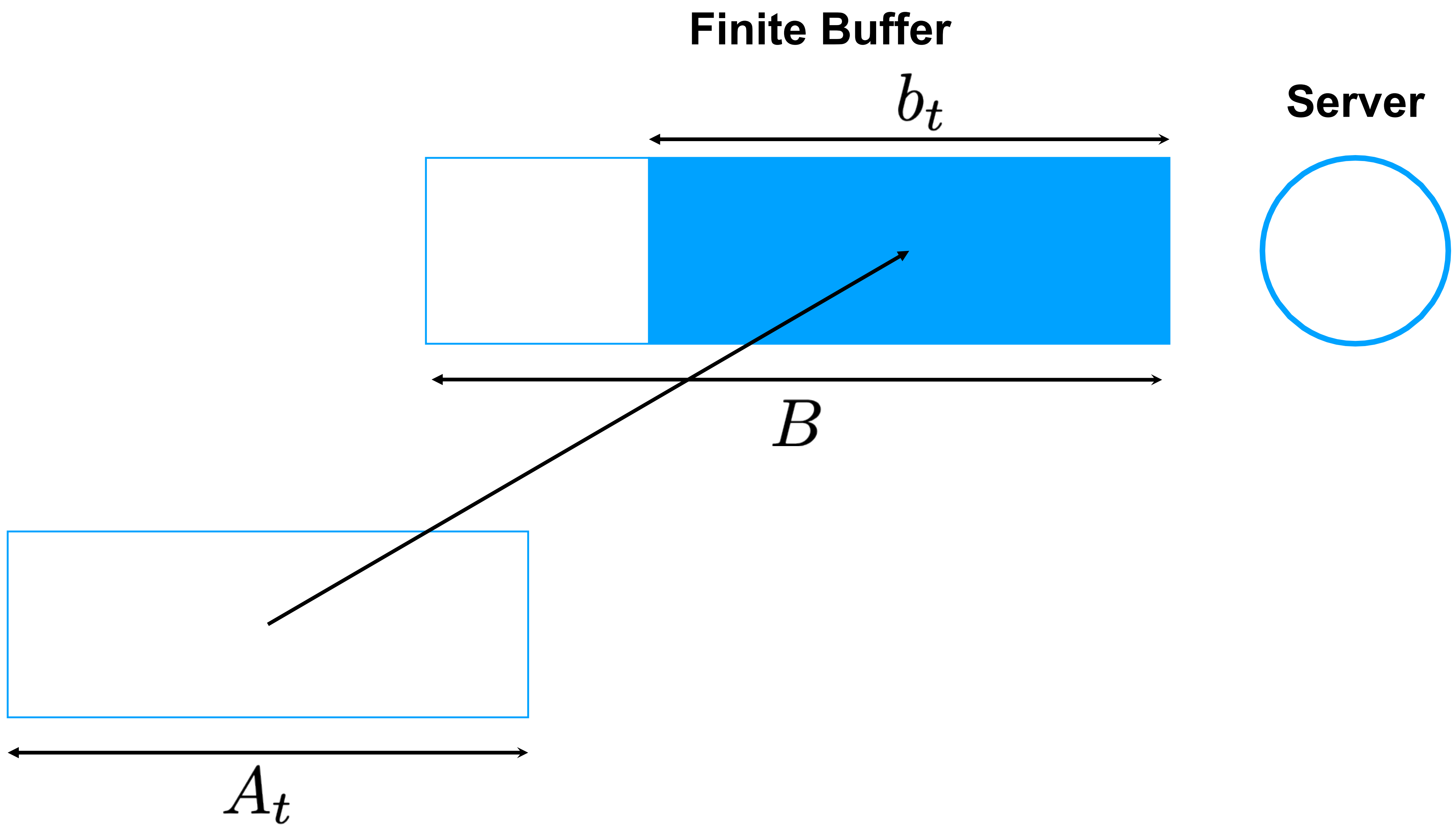
System Model



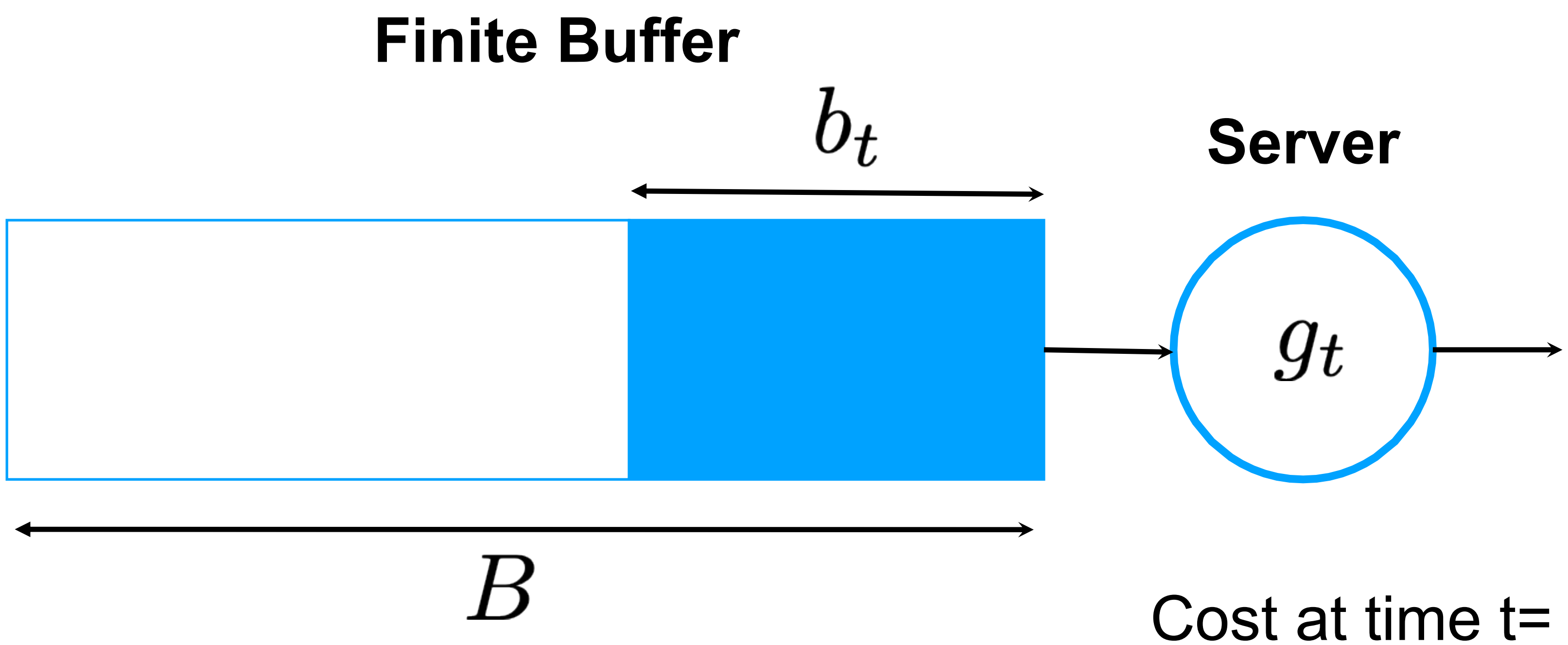
System Model



System Model



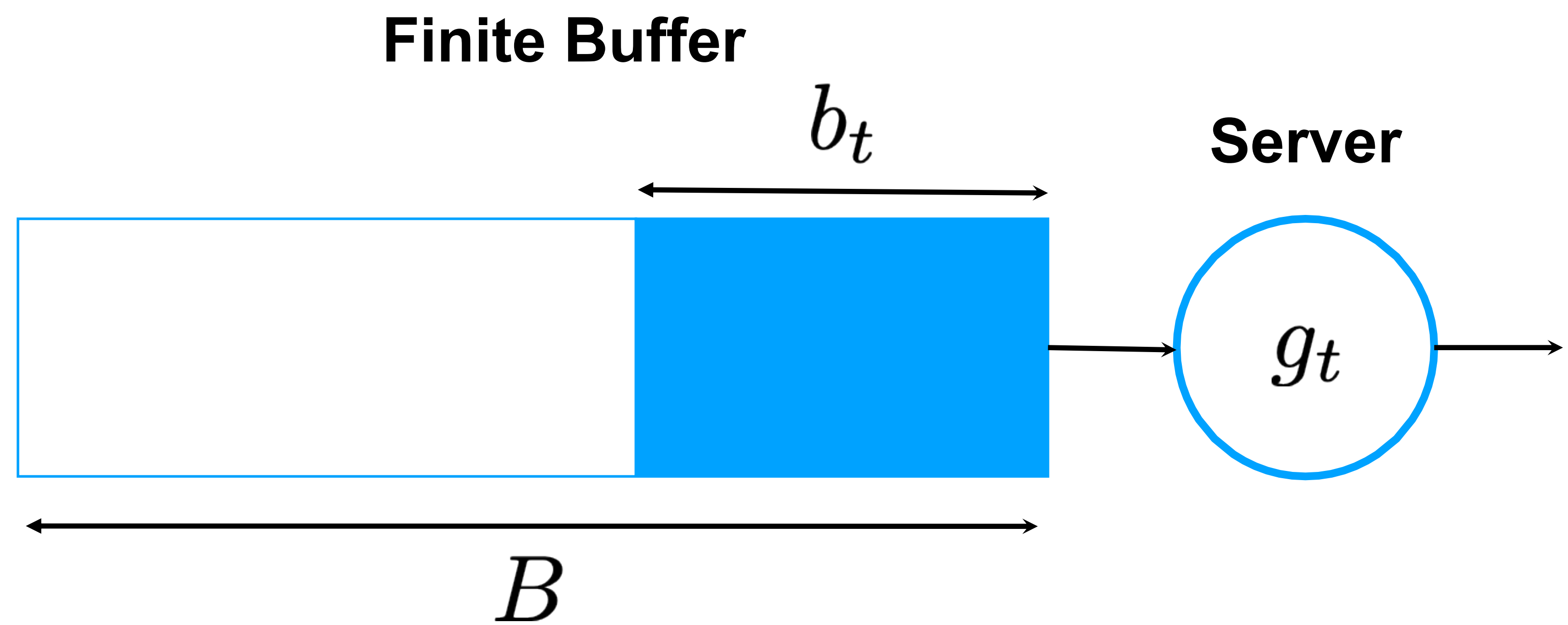
System Model

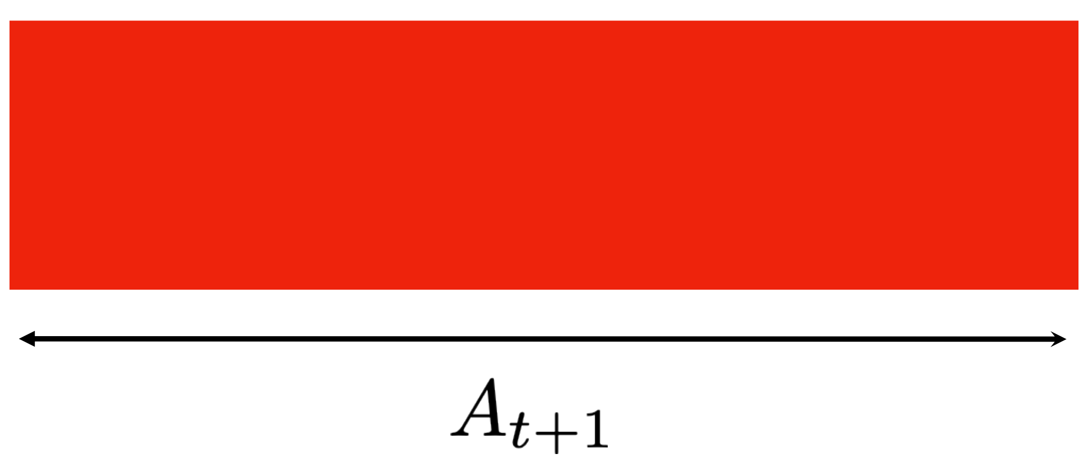


Cost at time t= g_t^2

Close approximation of microprocessor power cost (Wiermen et al. 2012)

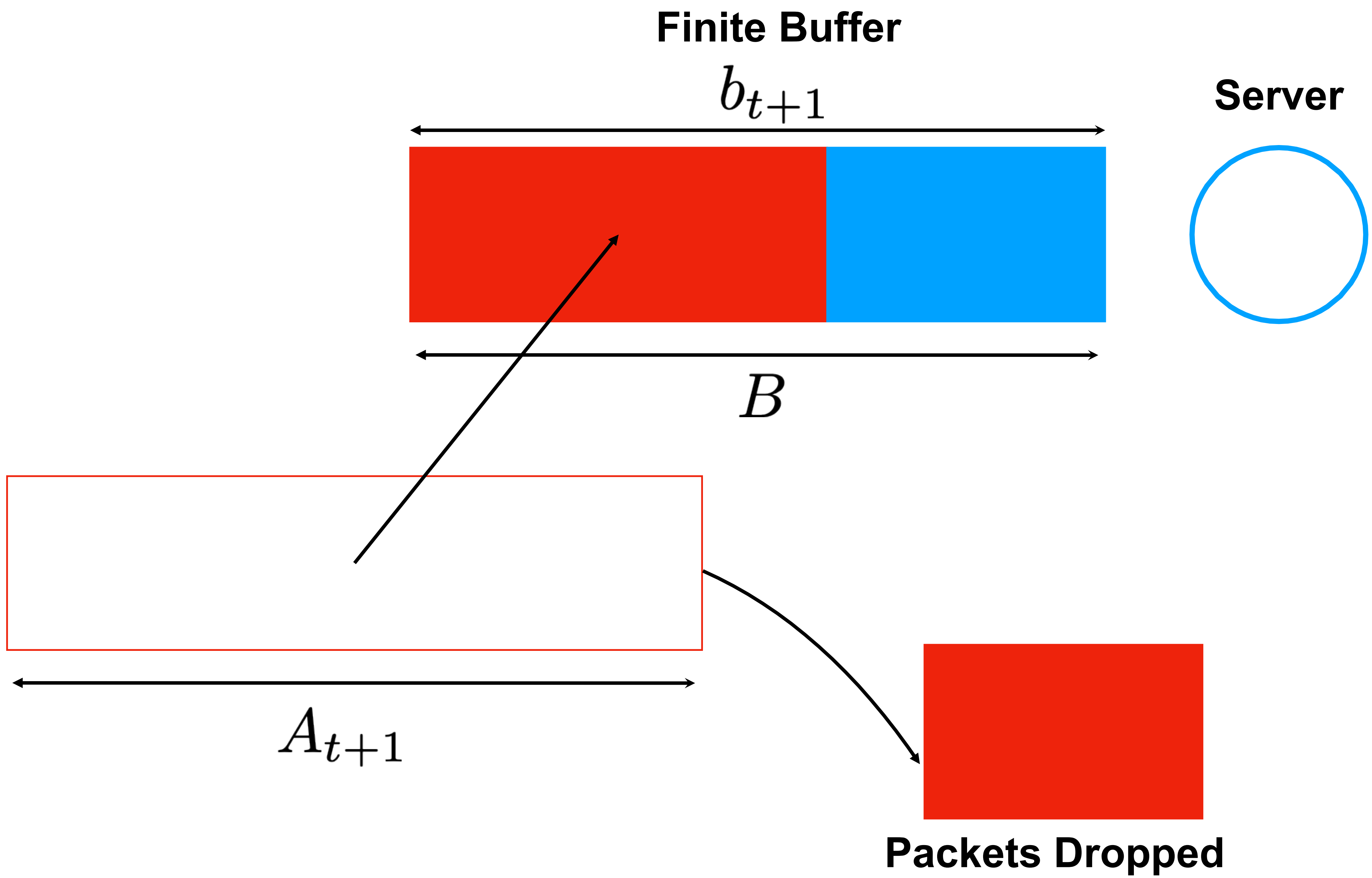
System Model



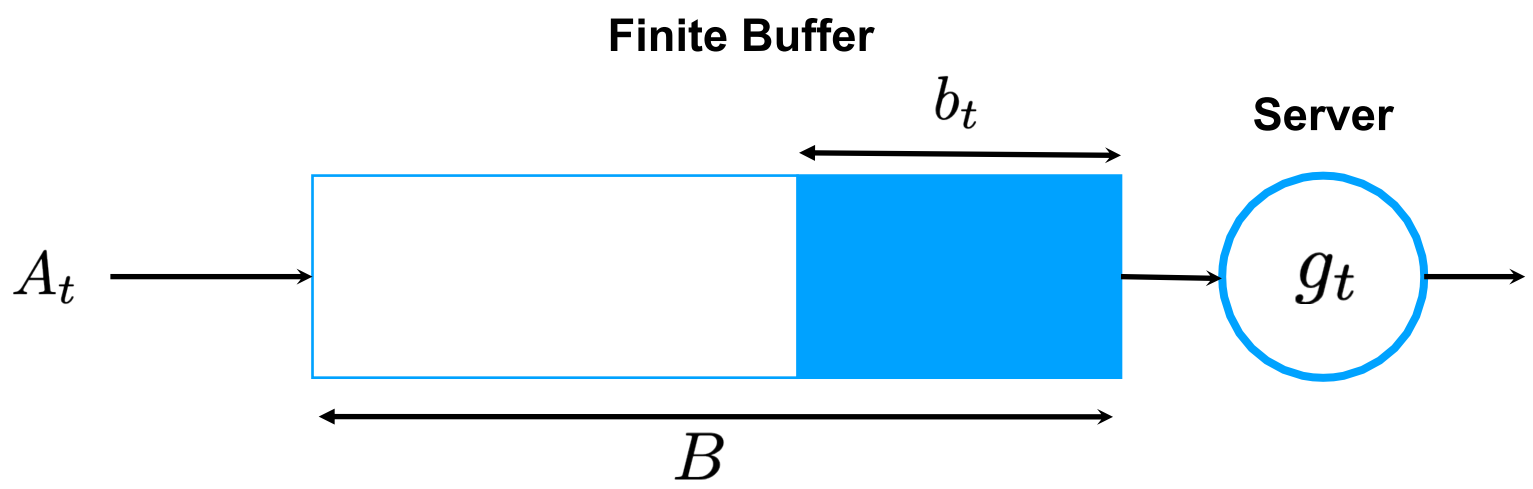


A_t, A_{t+1} **are**

System Model



Problem Formulation



- Overall Cost = $\frac{1}{n}\sum_{t=1}^{n} g_t^2$
- P_{drop} = (Total # of Packets Dropped)/(Total # of Packets Arrived)

- Constraint: $P_{\text{drop}} \leq \alpha$
- Objective: Minimise overall cost subject to constraint