Regenerative estimator of the overflow probability in a tandem network

Irina Dyudenko, Evsey Morozov, Michele Pagano

Institute of Applied Mathematical Research, Karelian Research Center and Petrozavodsk University

Dept. of Information Engineering, University of Pisa, Italy

Rennes – RESIM 2008

September 2008

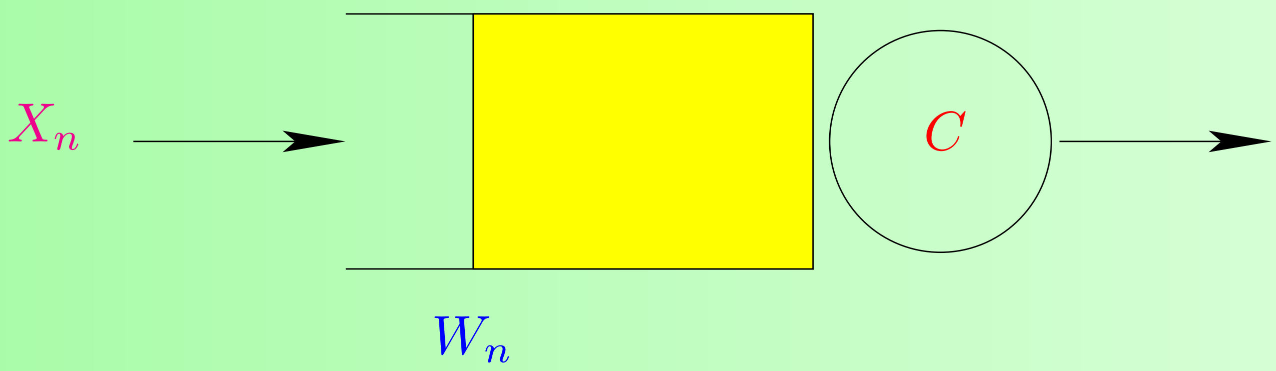
Outlines

- Network scenario
  - Quality of Service (QoS)
  - Bandwidth allocation techniques
- Effective Bandwidth (EB)
  - Definition and physical meaning
  - Scaled Cumulant Generating Function (SCGF)
- **Effective Bandwidth estimators**
  - Standard estimator: the batch-mean approach
  - Refined estimator for input traffic with a regenerative structure
- Performance Comparison
  - Tandem network with 2 queues and deterministic service rate
  - Variance of the estimator
- Conclusions

Network scenario

- Evolution of network services and architectures
  - Quality of Service (QoS) requirements
  - Packet loss (usually approximated through the overflow probability) as a typical QoS guarantee
  - Traffic Engineering
- Necessity of dynamic control mechanisms
  - Bandwidth allocation techniques
  - Admission Control
- The notion of effective bandwidth has emerged as a powerful metric to quantify the amount of resources needed by connections in order to guarantee a required QoS level
- Theorethical background: **Large Deviation Theory**
  - For a wide class of buffered systems the overflow probability decreases exponentially fast as buffer size increases
  - The rate-function allows to calculate the required exponent
- **Open issue: estimation of the effective bandwidth from traffic measurements**

Workload process in a single server queue



- Stationarity condition: $\mathbb{E}X_n < C$
- Workload process:

$$W_n = \sup_{u \le n} \left(\sum_{i=1}^{u} X_i - C \cdot u \right)$$

- Under mild assumptions, stationary workload $W_n \Rightarrow W$ exists and satisfies a Large Deviation Principle (LDP)

$$\mathbb{P}\,(W > b) \approx e^{-\delta\, b}$$

Large Deviation Principle for the workload process

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{P}(W > x) = -\delta$$

where

- the exponent is defined as

$$\delta = \delta(C) = \sup\{\theta > 0 : \Lambda(\theta) < C\theta\}$$

- the quantity

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} e^{\theta \sum_{i=1}^{n} X_i} \triangleq \lim_{n\to\infty} \Lambda_n(\theta)$$

  is known as Limiting Scaled Cumulant Generating Function (LSCGF)

In case of an i.i.d. sequence X_n, the previous limit leads to

$$\Lambda(\theta) = \log \mathbb{E} e^{\theta X}$$

Effective Bandwidth (EB)

- In real networks, it is reasonable to assume that the buffer size b is fixed, while the capacity C can be varied
- **Goal:** determine the minimum service capacity C_Γ which guarantees an overflow probability $\leq \Gamma$

$$C_\Gamma = \min(C : e^{-\delta(C)b} \leq \Gamma)$$

- Hence, we obtain

$$\delta(C_\Gamma) = \theta^* = -\frac{\log(\Gamma)}{b}$$

and, since

$$\delta = \delta(C) = \sup\{\theta > 0 : \Lambda(\theta) < C\theta\}$$

we can express C as a function of θ^* (provided that $\Lambda(\theta)$ is known):

$$C_\Gamma = C(\Gamma, b) = \frac{\Lambda(\theta^*)}{\theta^*}$$

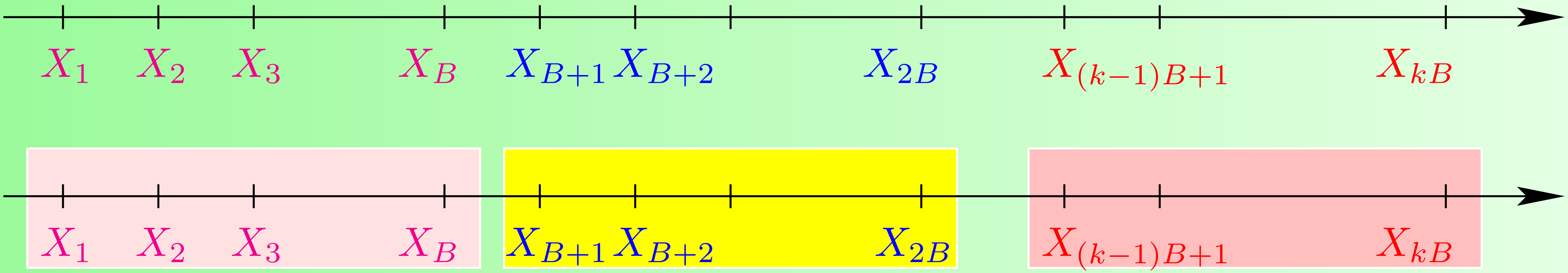
- Given the QoS parameter Γ and the buffer size b, the value $C_\Gamma = C(\Gamma, b)$ is called the effective bandwidth of the incoming traffic

EB Estimation: batch-mean approach

- If the samples X_i were i.i.d., then we could estimate the LSCGF for $\theta = \theta^*$ by means of the sample mean estimator

$$\hat{\Lambda}_n(\theta^*) \; = \; \log \frac{1}{n} \sum_{i=1}^{n} e^{\theta^* X_i} \; \rightarrow \; \log \mathbb{E}\left(e^{\theta^* X}\right)$$

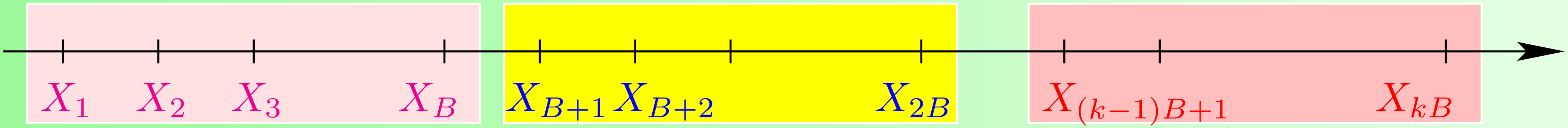
- Aggregation into blocks of size B



$$\hat{X}_1 = \sum_{i=1}^{B} X_i \qquad \hat{X}_2 = \sum_{i=B+1}^{2B} X_i \qquad \hat{X}_k = \sum_{i=(k-1)B+1}^{kB} X_i$$

- If B is large enough, the dependence between blocks *should be negligible*, and $\hat{X}_i$ constitute *(approximately) an i.i.d. sequence*

EB Estimation: batch-mean approach



$$\hat{X}_1 = \sum_{i=1}^{B} X_i \qquad \hat{X}_2 = \sum_{i=B+1}^{2B} X_i \qquad \hat{X}_k = \sum_{i=(k-1)B+1}^{kB} X_i$$

- By using $n = kB$ observations, we obtain a modified expression of the LSCGF

$$\Lambda(\theta^*, B) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} e^{\theta^* \sum_{i=1}^{k} \hat{X}_i} = \lim_{n\to\infty} \frac{k}{n} \log \mathbb{E} e^{\theta^* \hat{X}} = \frac{1}{B} \log \mathbb{E} e^{\theta^* \hat{X}}$$

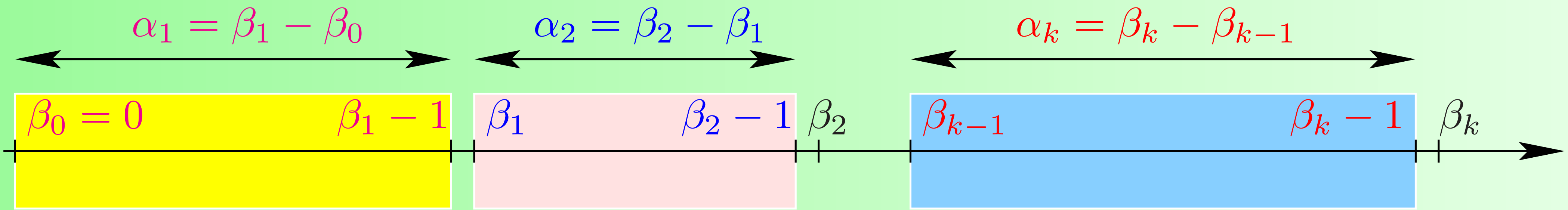
- Batch-mean estimator for the LSCGF

$$\hat{\Lambda}_n(\theta^*, B) = \frac{1}{B} \log \frac{1}{k} \sum_{i=1}^{k} e^{\theta^* \hat{X}_i} = \frac{1}{B} \log \frac{B}{n} \sum_{i=1}^{n/B} e^{\theta^* \hat{X}_i}$$

- Batch-mean estimator for the EB

$$\hat{C}(\Gamma, b) = \frac{\hat{\Lambda}_n(\theta^*, B)}{\theta^*}$$

EB Estimation: regenerative approach



$$\hat{X}_1 = \sum_{i=\beta_0}^{\beta_1 - 1} X_i \qquad \hat{X}_1 = \sum_{i=\beta_1}^{\beta_2 - 1} X_i \qquad \hat{X}_k = \sum_{i=\beta_{k-1}}^{\beta_k - 1} X_i$$

- **Assumption:** regenerative structure of the input with regeneration points β_n
  - randomization of the block size with $\mathbb{E}\alpha < \infty$
  - RVs $\hat{X}_n$ are i.i.d.
- Alternative definition of the EB estimator (which uses k regenerative blocks, corresponding to β_k observations)

$$\hat{\Lambda}_k(\theta^*) = \frac{k}{\beta_k} \log \frac{1}{k} \sum_{i=1}^{k} e^{\theta^* \hat{X}_i}$$

Heuristic justification

☞ Considering a large number of time slots n including k regeneration cycles, i.e.

$$n \;=\; \beta_k + (n - \beta_k) \qquad \text{where} \quad \frac{n-\beta_k}{n} \to 0$$

☞ Roughly speaking, we have:

$$\frac{1}{n}\log \mathbb{E}e^{\theta^* \sum_{i=1}^{n} X_i} \approx \frac{1}{\beta_k}\log \mathbb{E}e^{\theta^* \sum_{i=1}^{k} \hat{X}_i} = \frac{k}{\beta_k}\log \mathbb{E}e^{\theta^* \hat{X}}$$

☞ From renewal theory

$$\frac{k}{\beta_k} \;\to\; \frac{1}{\mathbb{E}\alpha}$$

☞ By the Strong Law of Large Numbers

$$\frac{1}{k}\sum_{i=1}^{k} e^{\theta^* \hat{X}_i} \to \mathbb{E}e^{\theta^* \hat{X}}$$

☞ Hence

$$\hat{\Lambda}_k(\theta^*) = \frac{k}{\beta_k}\log \frac{1}{k}\sum_{i=1}^{k} e^{\theta^* \hat{X}_i} \;\to\; \frac{1}{\mathbb{E}\alpha}\log \mathbb{E}e^{\theta^* \hat{X}}$$

Lower bound

- Notation:
  - $k(n) \triangleq \max(k : \beta_k \leq n)$
  - $\mathbb{E}e^{\theta^* \hat{X}} \triangleq a < \infty$
- By using conditional expectation and taking into account the independence between $k(n)$ and the regenerative blocks

$$\frac{1}{n}\log \mathbb{E}e^{\theta^* \sum_{i=1}^{n} X_i} \geq \frac{1}{n}\log \mathbb{E}\left(\mathbb{E}e^{\theta^* \sum_{i=1}^{k(n)} \hat{X}_i} | k(n)\right) = \frac{1}{n}\log \mathbb{E}a^{k(n)}$$

- By Jensen's inequality

$$\log \mathbb{E}a^{k(n)} \geq \mathbb{E}\log a^{k(n)} = \mathbb{E}k(n)\log a$$

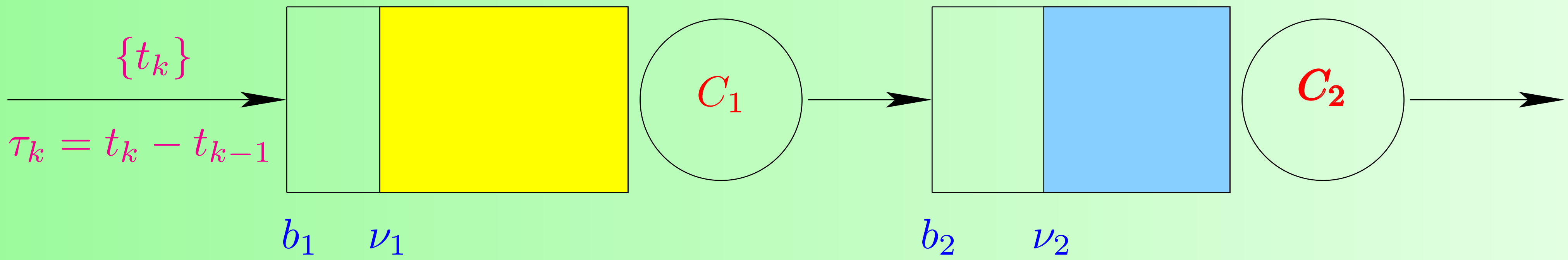
- From the elementary renewal theorem

$$\frac{\mathbb{E}k(n)}{n} \rightarrow \frac{1}{\mathbb{E}\alpha}$$

- Hence we obtain the desired lower bound:

$$\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{E}e^{\theta^* \sum_{i=1}^{n} X_i} \geq \frac{\log a}{\mathbb{E}\alpha} \triangleq \frac{1}{\mathbb{E}\alpha} \log \mathbb{E}e^{\theta^* \hat{X}}$$

Simulation Scenario



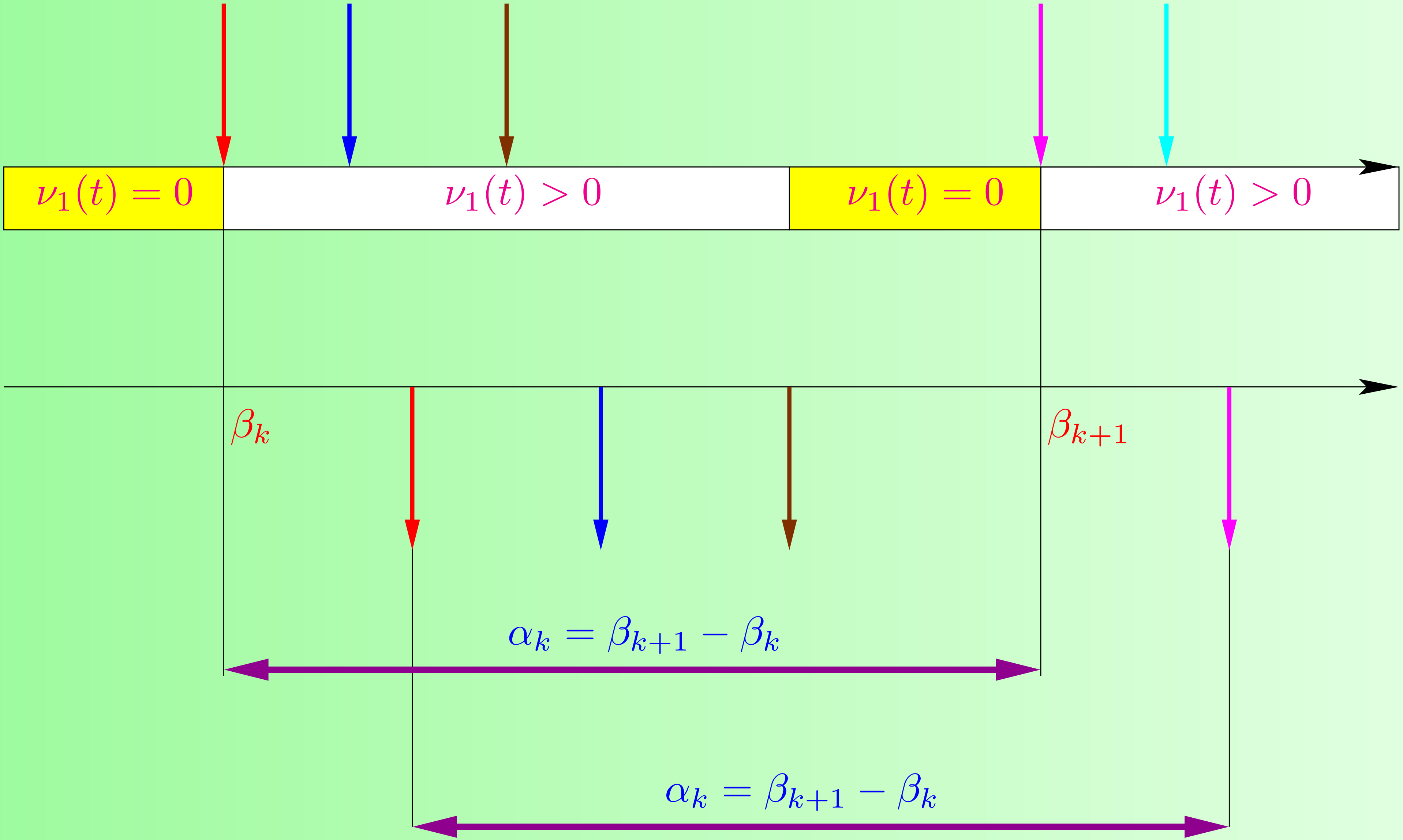
- The first station is fed by a renewal input
- *Minimal* construction of regenerations for the input to the second station

$$\begin{cases} \beta_0 & = 0 \\ \beta_{n+1} & = \min_k \left(t_k > \beta_n : \nu_1(t_k^-) = 0 \right) \qquad n \geq 0 \end{cases}$$

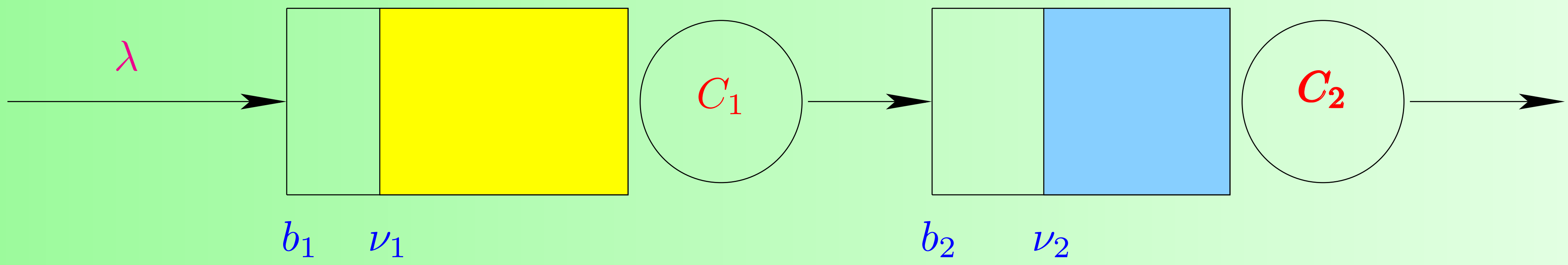
- Since the queue-size is bounded, if $\mathbb{P}\left(\tau > 1/C_1\right) > 0$ then the renewal process $\beta = (\beta_n)$ is positive recurrent, that is it has a finite mean $\mathbb{E}\alpha < \infty$
- The target parameter to be estimated is the (constant) service rate C_2 for a given QoS level (overflow probability $\leq \Gamma$)

Regeneration Cycle

Packet arrivals and departure at the first queue



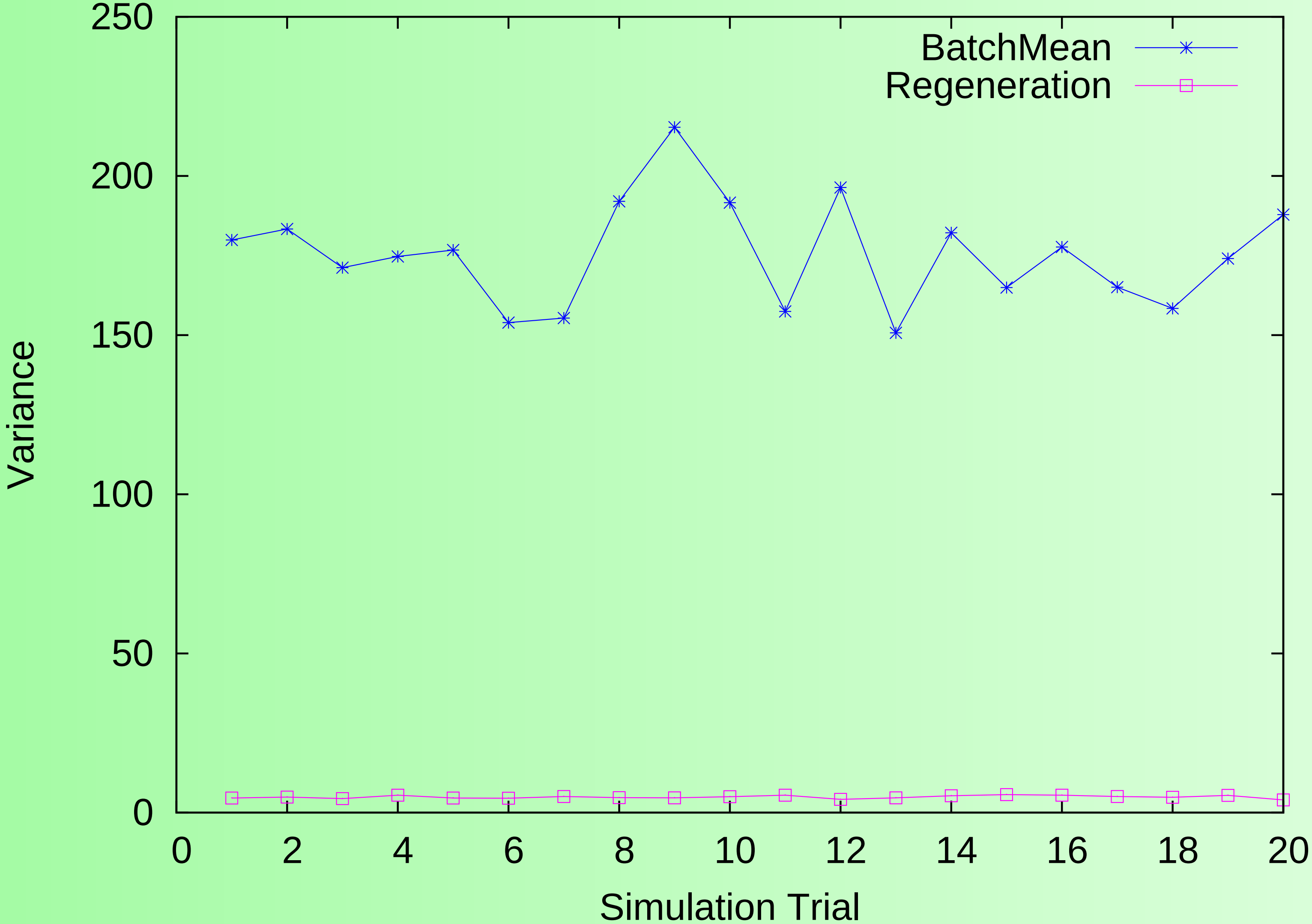
Simulation set–up



- The input traffic is *Poisson*(λ), i.e. $\tau \in \mathcal{E}\left(\lambda\right)$
- We compare two different estimatators
  - Regenerative approach ($\hat{C}_2^{\text{REG}}$)
  - Batch-mean approach ($\hat{C}_2^{\text{BM}}$)
    - The block size B is chosen according to $\mathcal{U}\left[500;\ 20000\right]$
    - For a given value of B, 1000 blocks are considered
    - 200 estimates are generated for different values of B
- Then variances of both estimators are calculated based on 200 *sample paths*

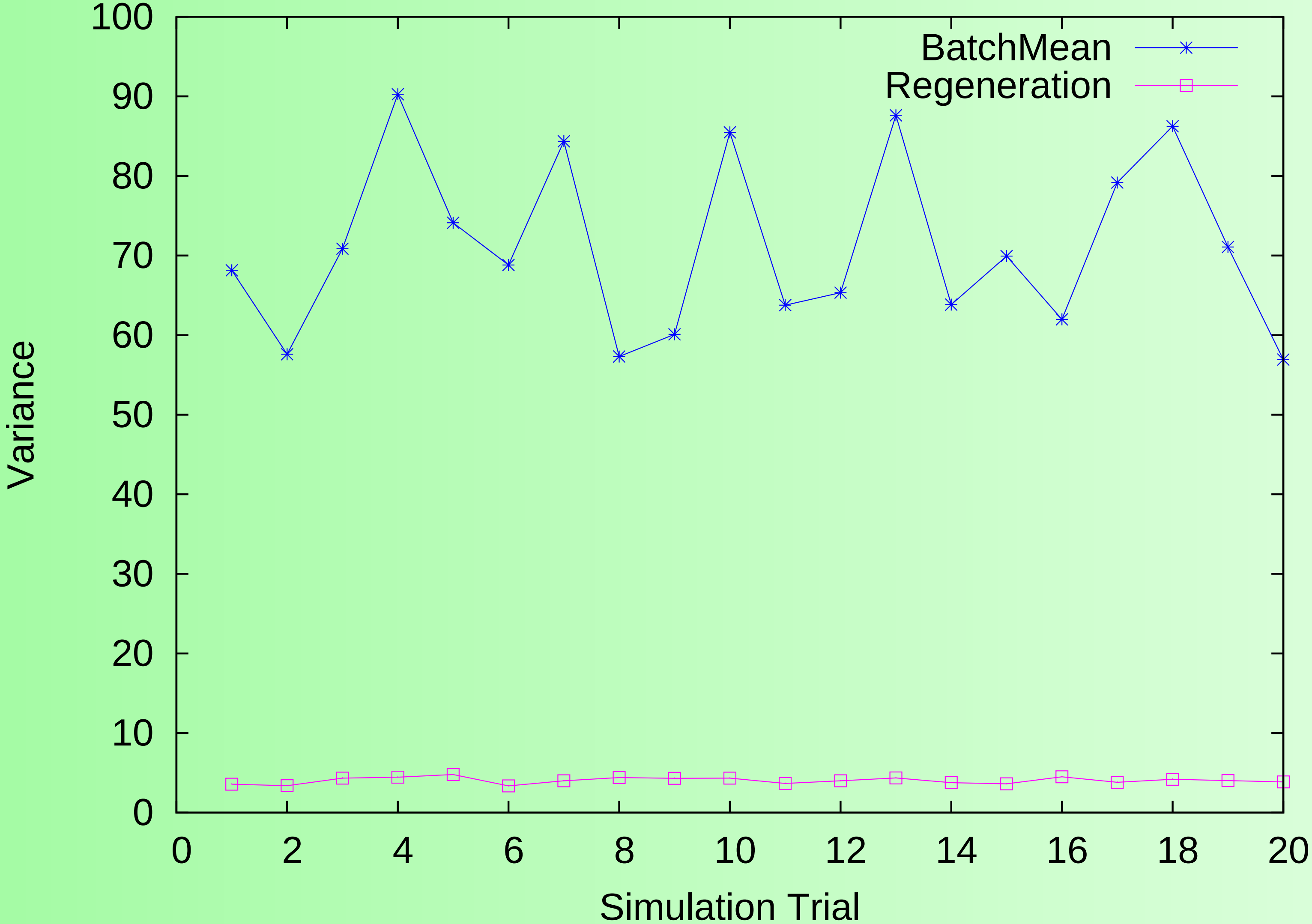
Simulation Results - $b_2 = 60$

Simulation Parameters: $C_1 = 0.35, \lambda = 0.2, b_1 = 100, \Gamma = 0.0001$



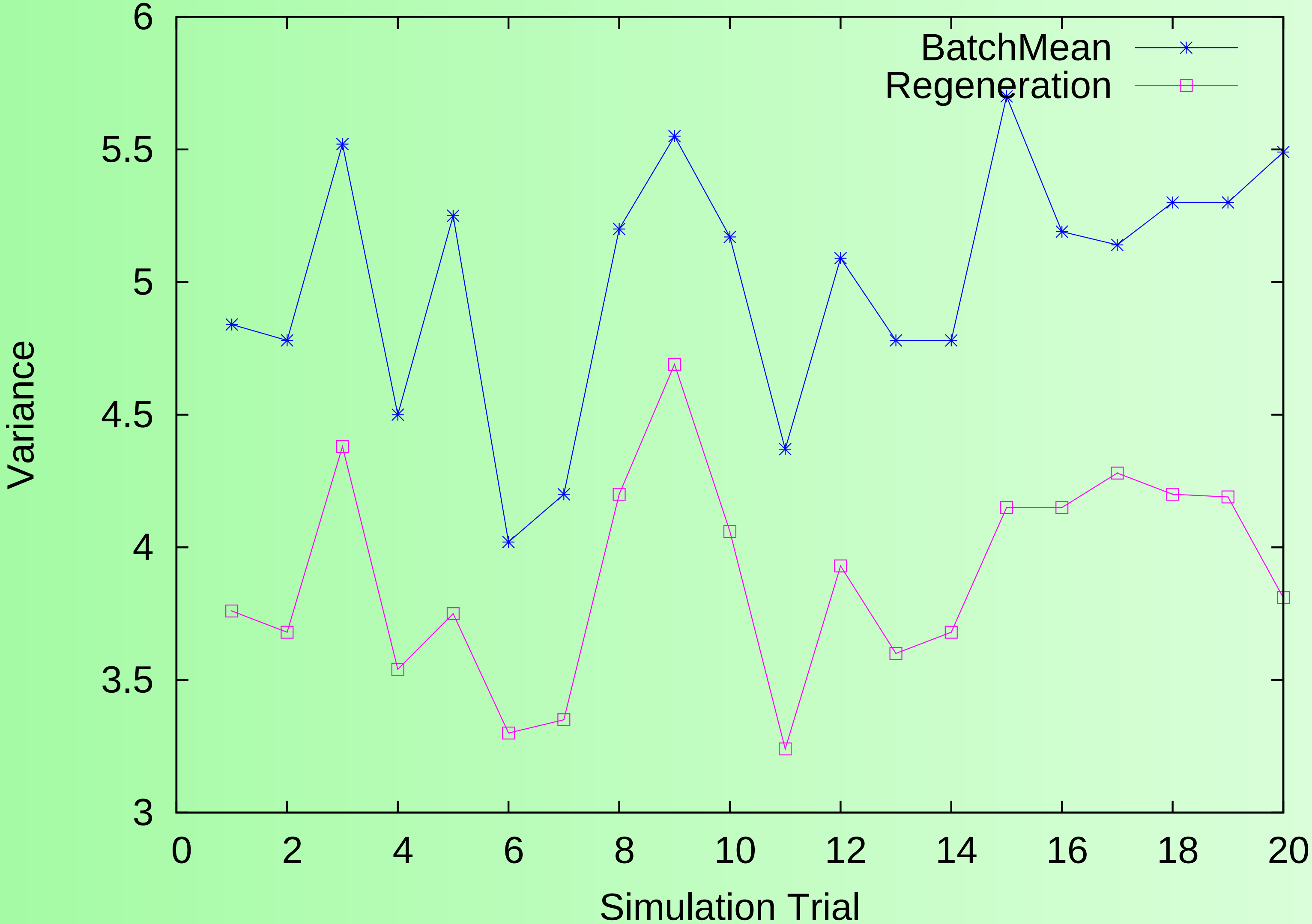
Simulation Results - $b_2 = 100$

Simulation Parameters: $C_1 = 0.35, \lambda = 0.2, b_1 = 100, \Gamma = 0.0001$



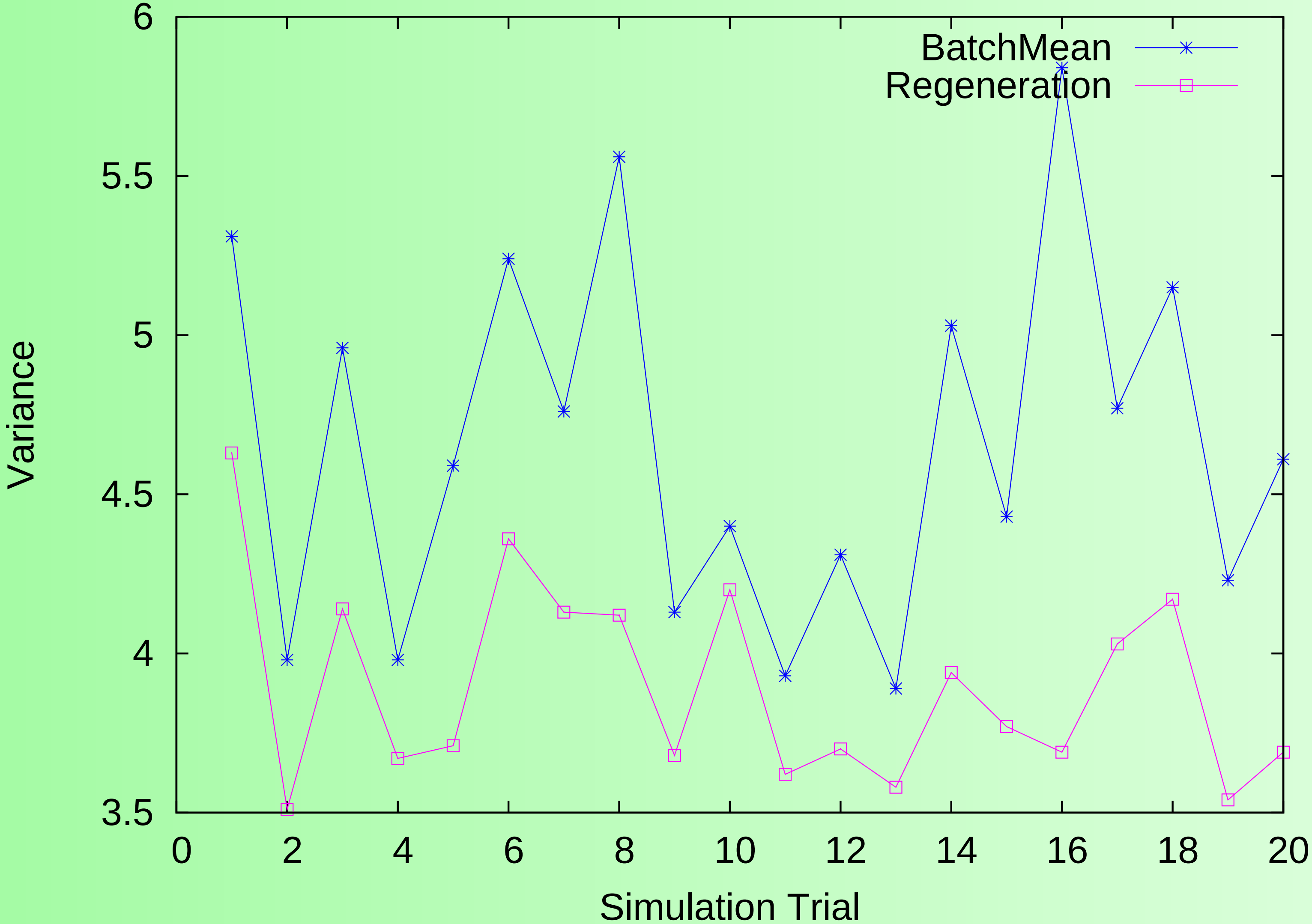
Simulation Results - $b_2 = 1000$

Simulation Parameters: $C_1 = 0.35, \lambda = 0.2, b_1 = 100, \Gamma = 0.0001$



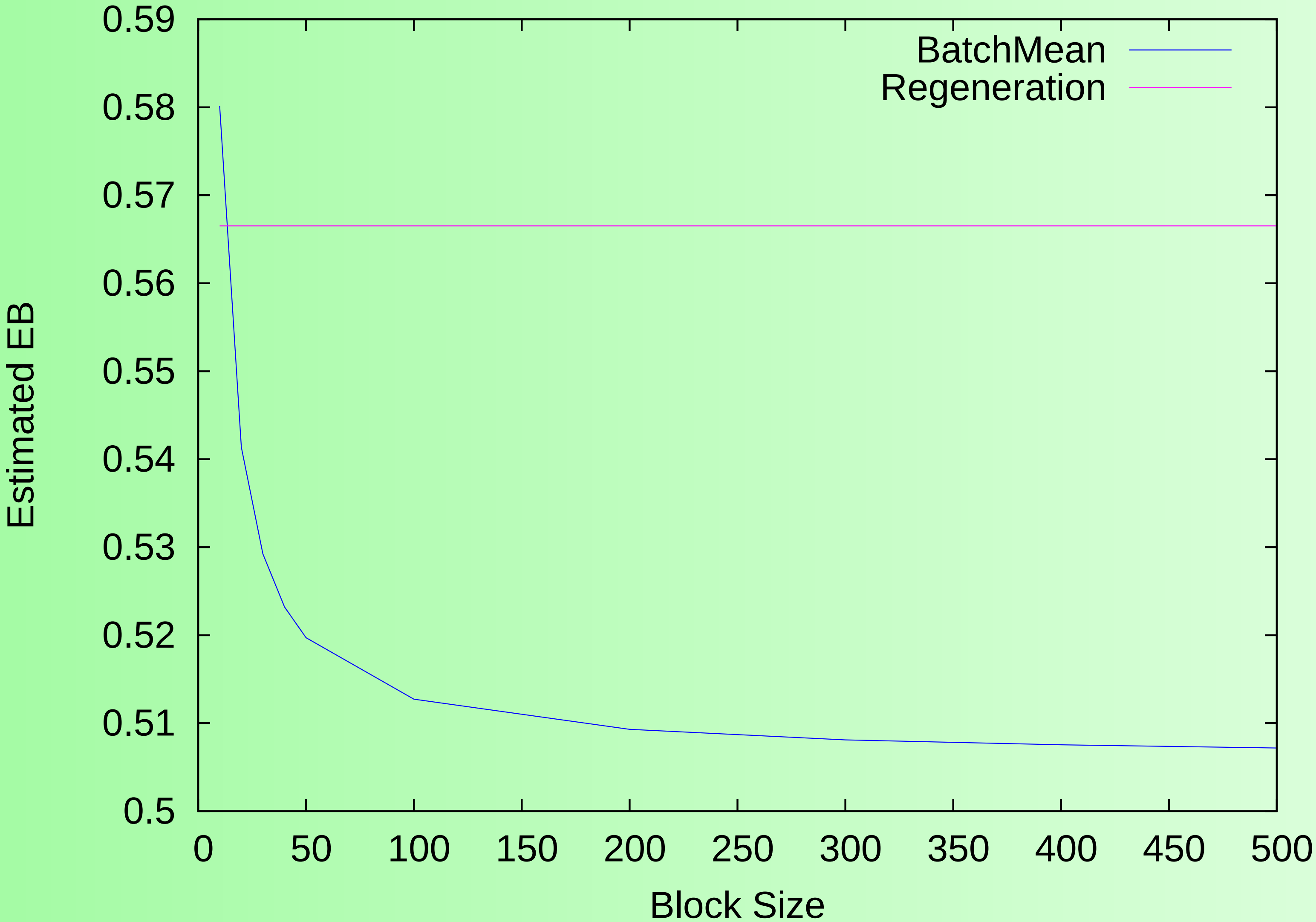
Simulation Results - $b_2 = 10000$

Simulation Parameters: $C_1 = 0.35, \lambda = 0.2, b_1 = 100, \Gamma = 0.0001$



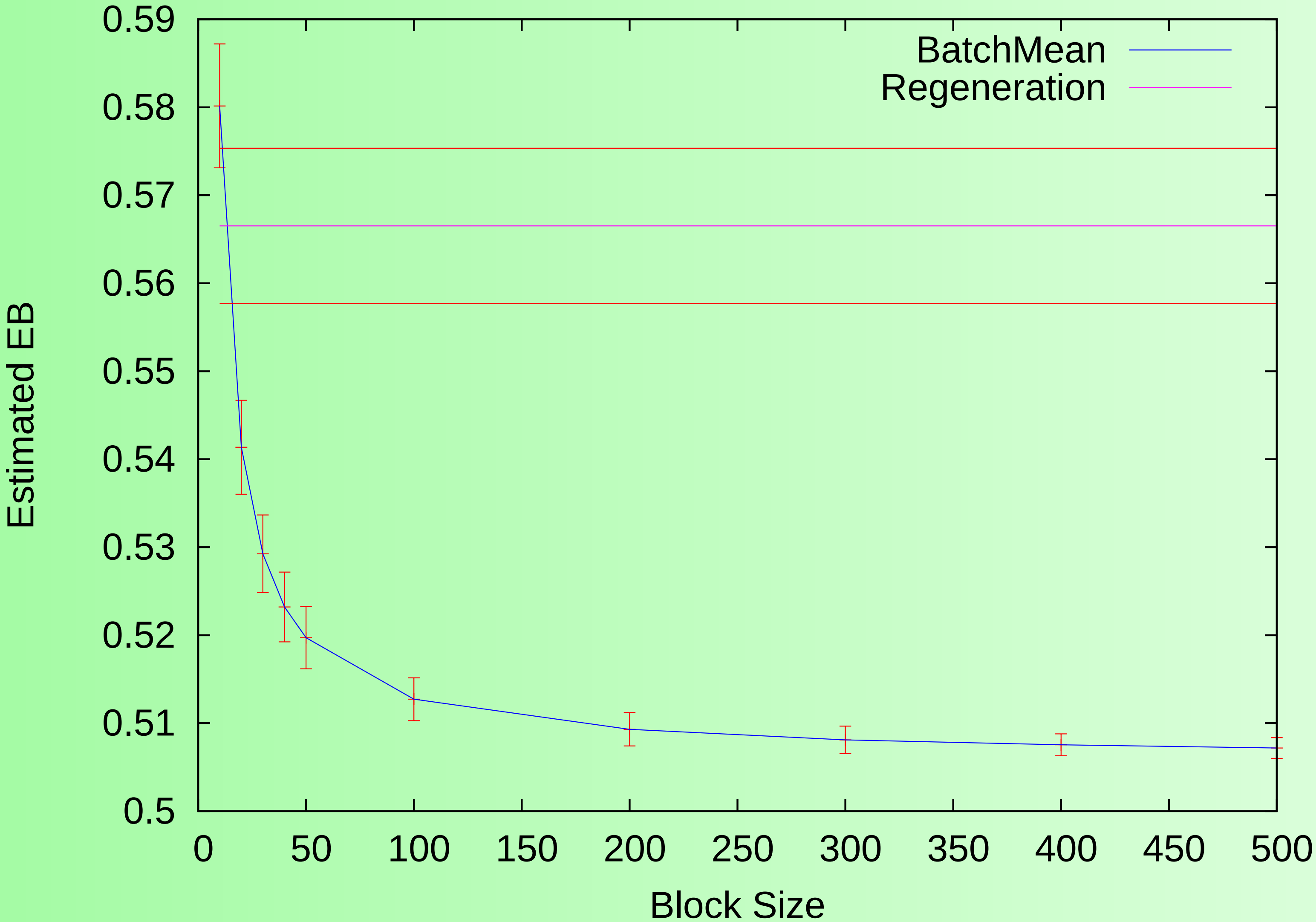
Simulation Results - block size

Simulation Parameters: $C_1 = 0.7, \lambda = 0.5, b_1 = 100, b_2 = 500, \Gamma = 0.00001$



Simulation Results - block size

Simulation Parameters: $C_1 = 0.7, \lambda = 0.5, b_1 = 100, b_2 = 500, \Gamma = 0.00001$



Conclusions

- This work deals with Effective Bandwidths of QoS demanding traffic flows
- In particular a refined version of the mean bach EB estimator is proposed
- The novelty of the contribution relies on the randomization of the block size estimator based on the identification of suitable renewall cycles
- Experimental results
  - The estimator has been tested in a simple tandem network topology
  - Simulation results show that the new estimator significantly outperforms the traditional batch mean approach in terms of estimation variance
- Work in progress ...
  - Simulations will be extended to a wider class of network topologies as well as input processes
  - Upper bound (to prove the convergence of the estimator)