Load Balancing Using Limited State Information

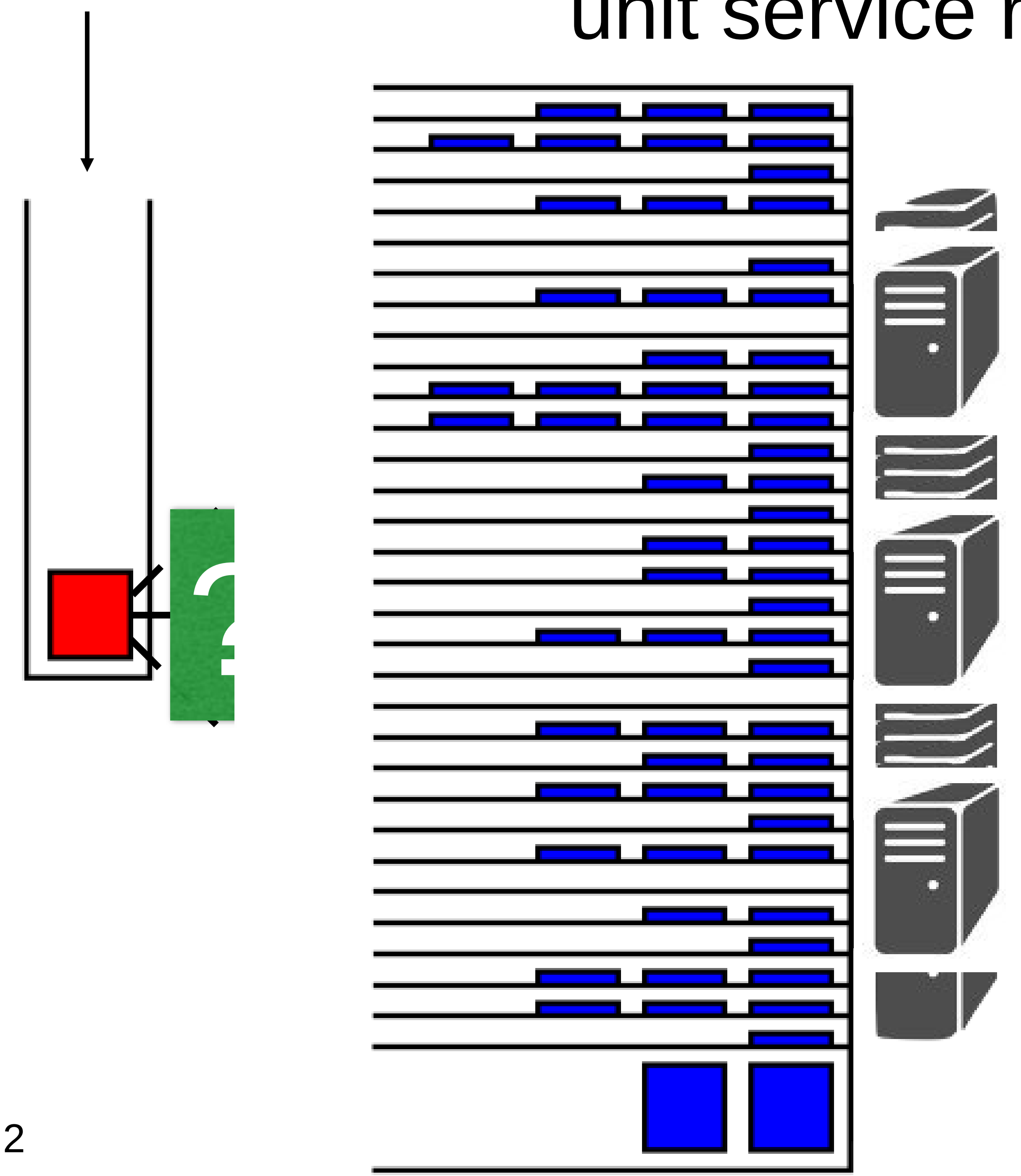
R. Srikant*
Joint work with Lei Ying+ and Xiaohan Kang+

*University of Illinois at Urbana-Champaign
+Arizona State University

Load Balancing

- Arriving tasks have to be routed to a server
- Requirement: small delays
- Join-the-shortest-queue
- Expensive feedback overhead

n servers with unit service rate

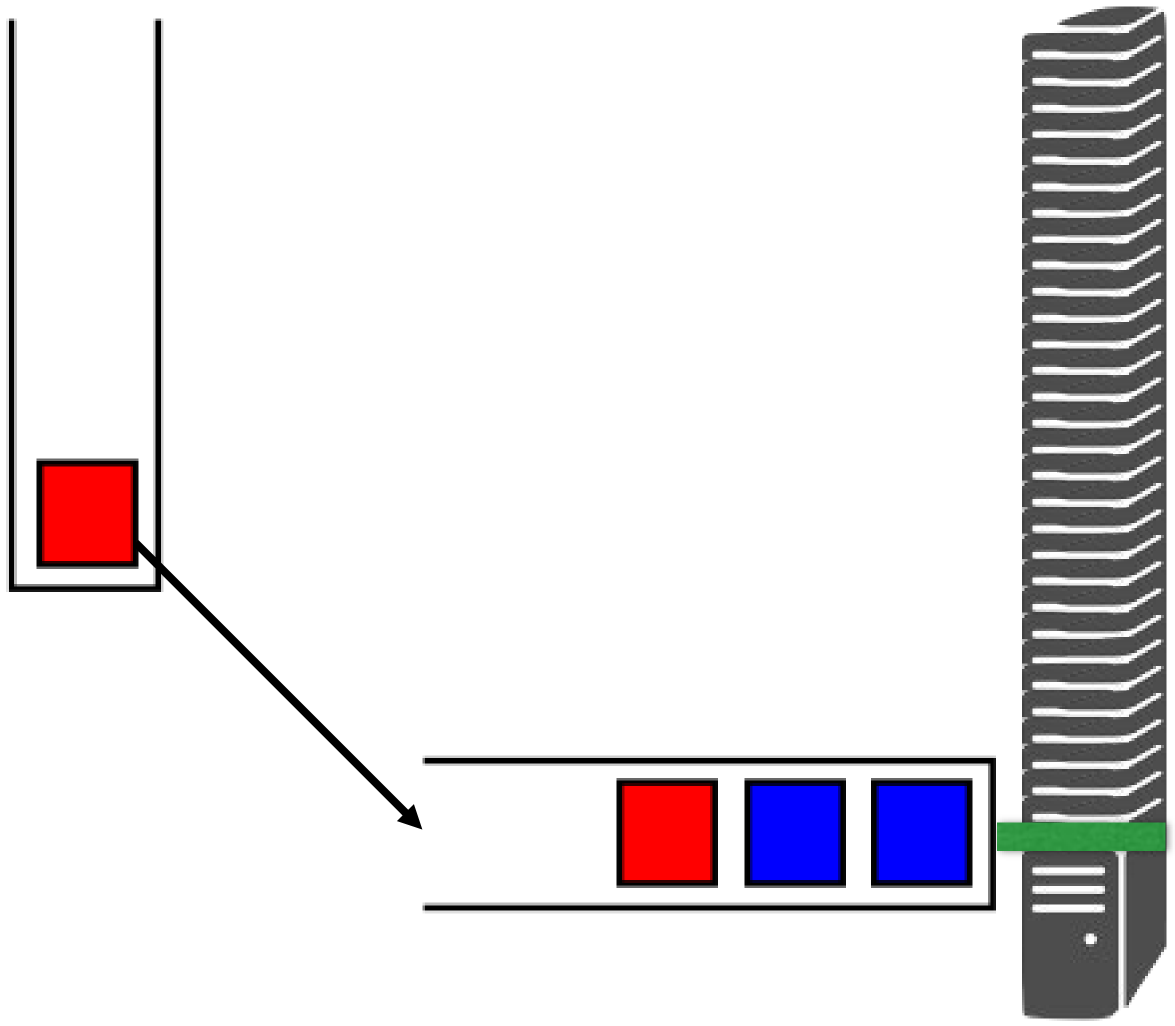


Data Centers are Large

Yahoo! Hadoop cluster
42,000 nodes

Random Routing

- No overhead
- Delay ~ $\dfrac{1}{1-\rho}$
- ρ: Traffic Intensity, i.e., the ratio of the arrival rate of tasks to the maximum rate at which they can be processed by the servers

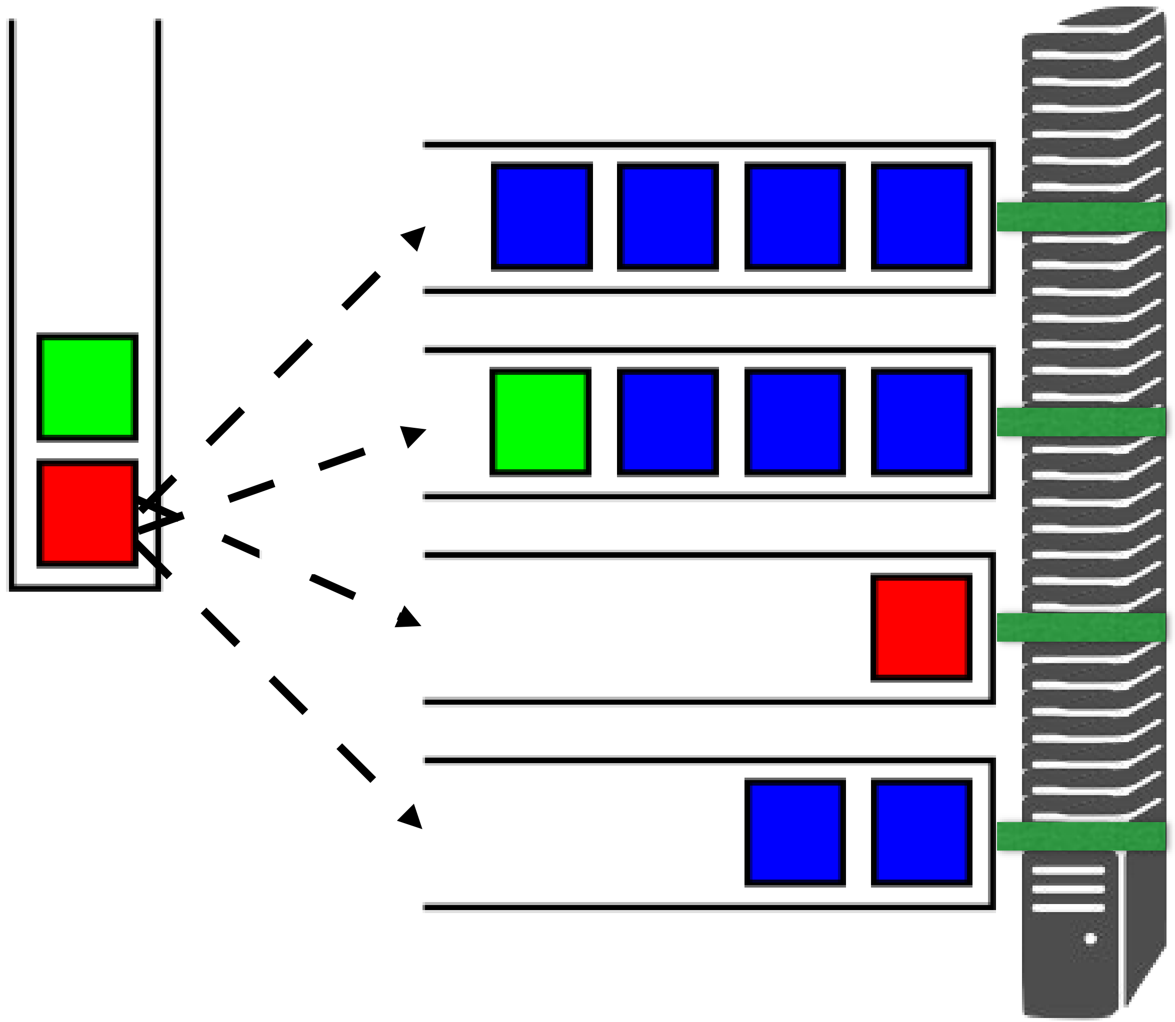


Power-of-Two-Choices

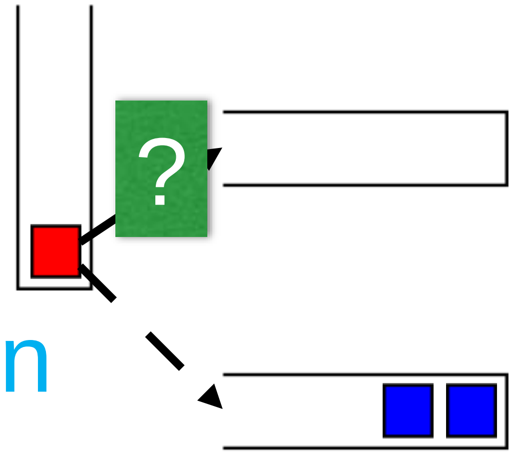
- Mitzenmacher 1996,
- Vvedenskaya, Dobrushin & Karpelevich 1996
- Delay, many-servers limit

$$\sum_{i=1}^{\infty} \rho^{2^i-2}$$

$$\rightarrow \boxed{\log_2 \frac{1}{1-\rho}} \qquad (\rho \approx 1)$$

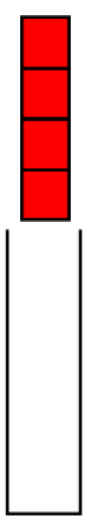


Key Question



- Question: sample one queue per arrival, on average, instead of two?

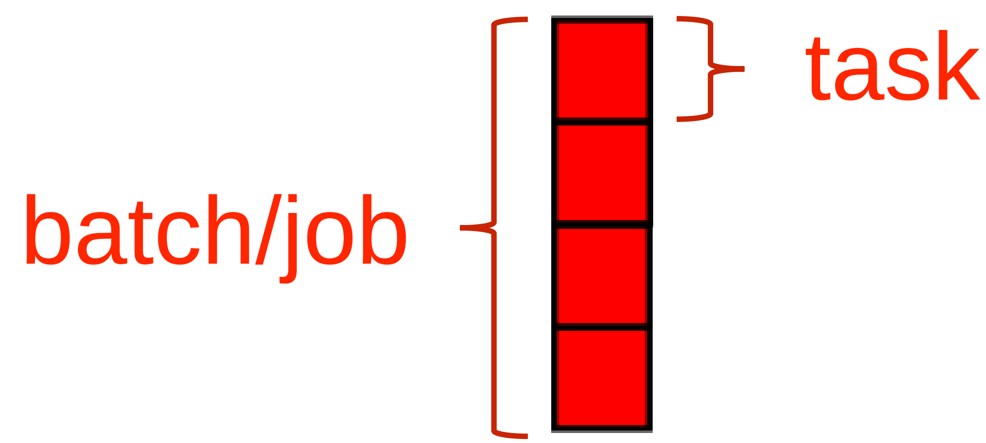
- Observation (Ousterhout et al, 2013): job arrivals occur in large batches (parallel tasks)



Model and Main Results

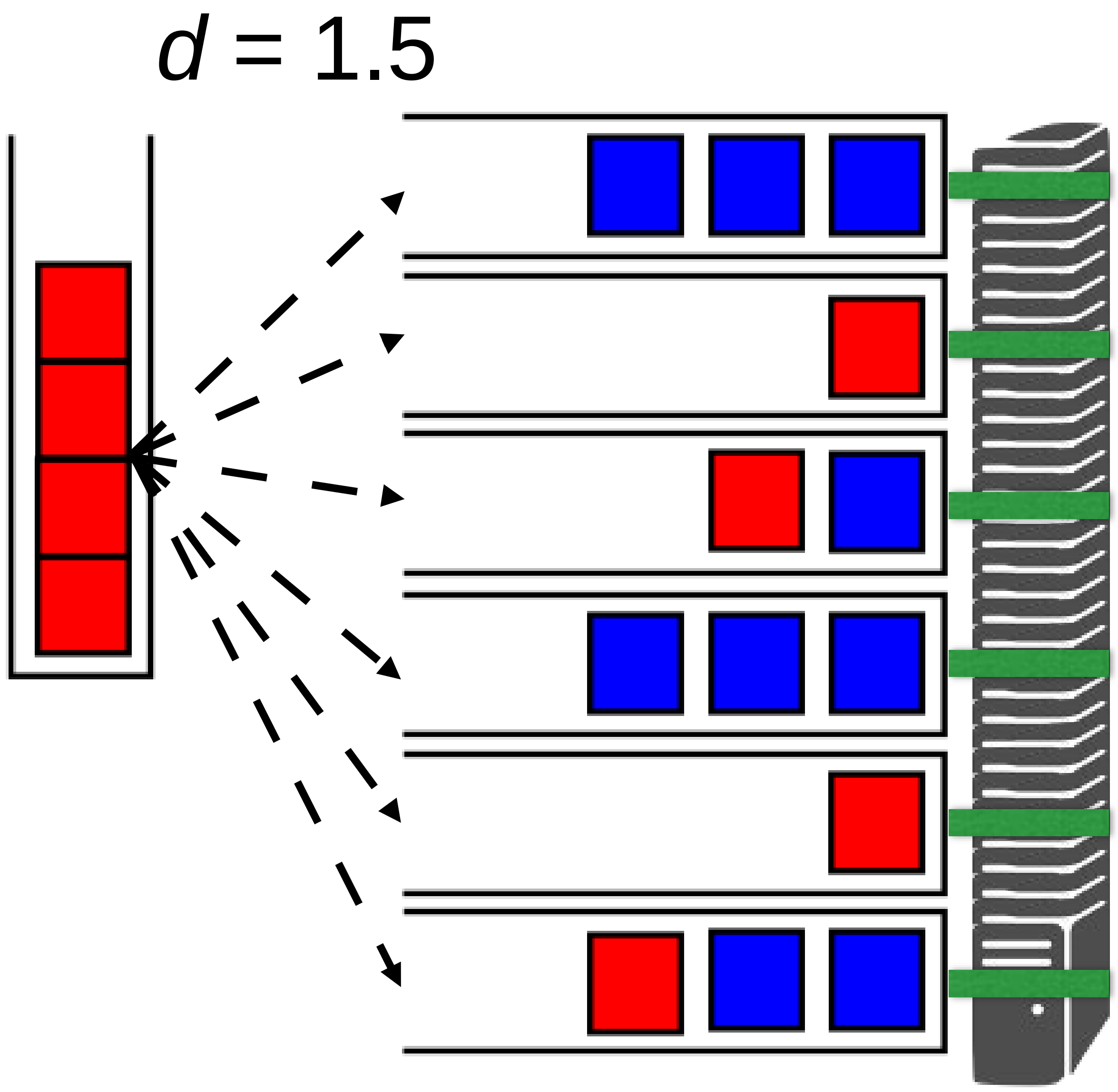
Model

- n servers
- Fixed batch size m $\boxed{1 \ll m \ll n}$
- Poisson batch arrivals with rate $n\rho/m$
- Note: Job is completed when all tasks in the job are completed
- Exponentially distributed service



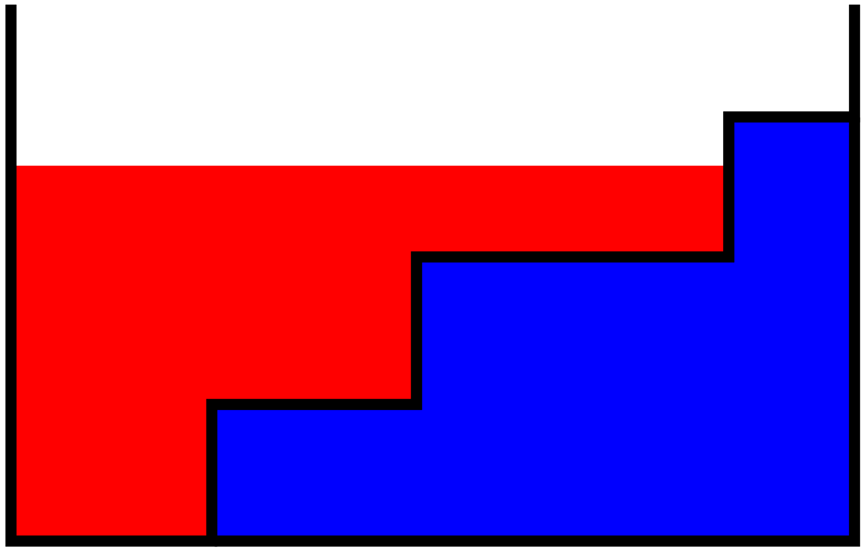
Batch-Sampling (BatchSamp)

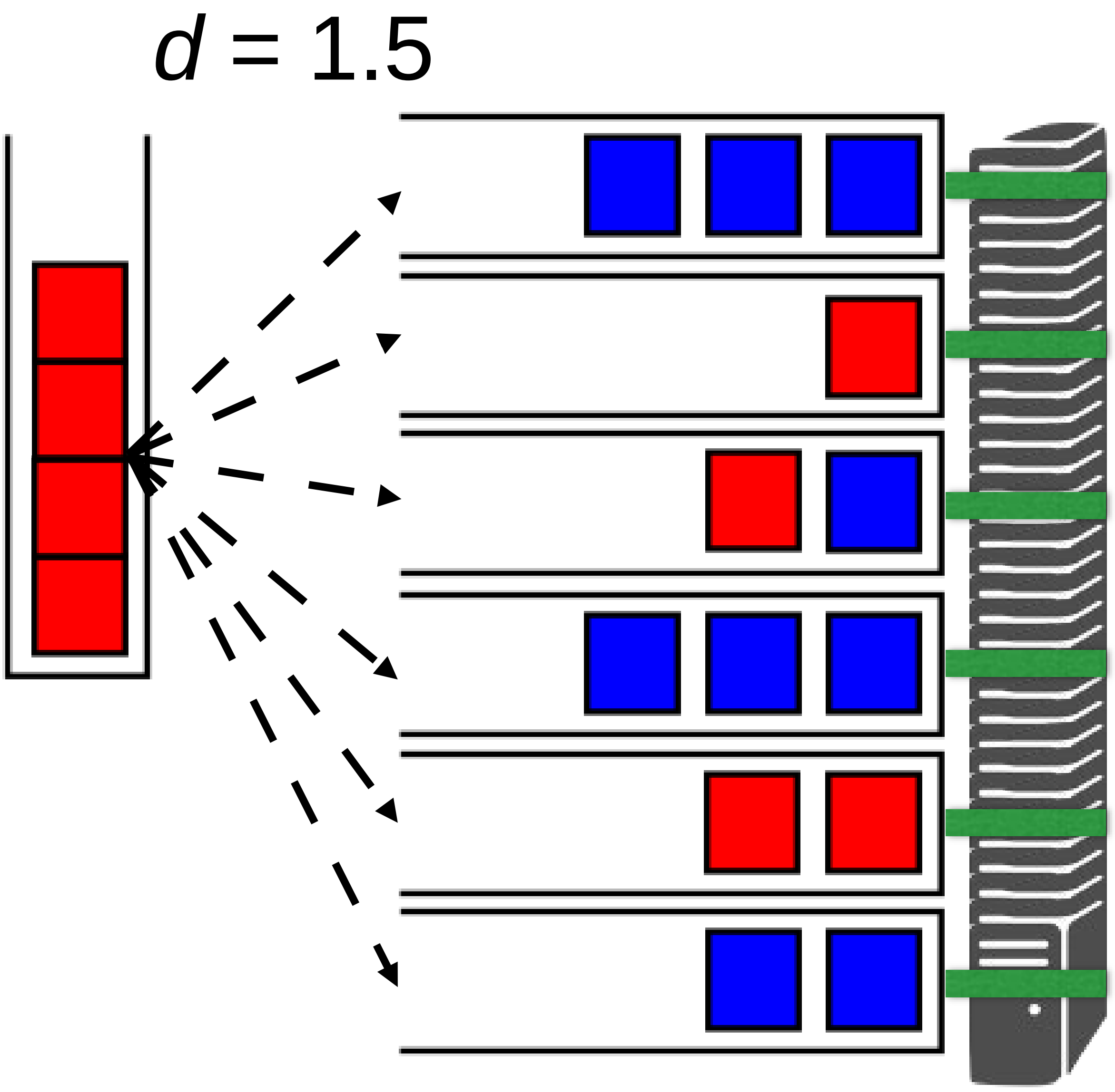
- Ousterhout et al. 2013
- Probe ratio *d*
- One job in each of the smallest queues



Batch-Filling (BatchFill)

Our algorithm:
Batch Sampling
+
WaterFilling





BatchSamp vs BatchFill

- Suppose we sample six queues to distribute five tasks: let's say that the queue lengths of the sampled queues are 0, 1, 2, 15, 20, 25
- Under BatchSamp, the resulting queue lengths are 1, 2, 3, 16, 21, 25
- Under BatchFill, they are 2, 3, 3, 15, 20, 25

BatchFill

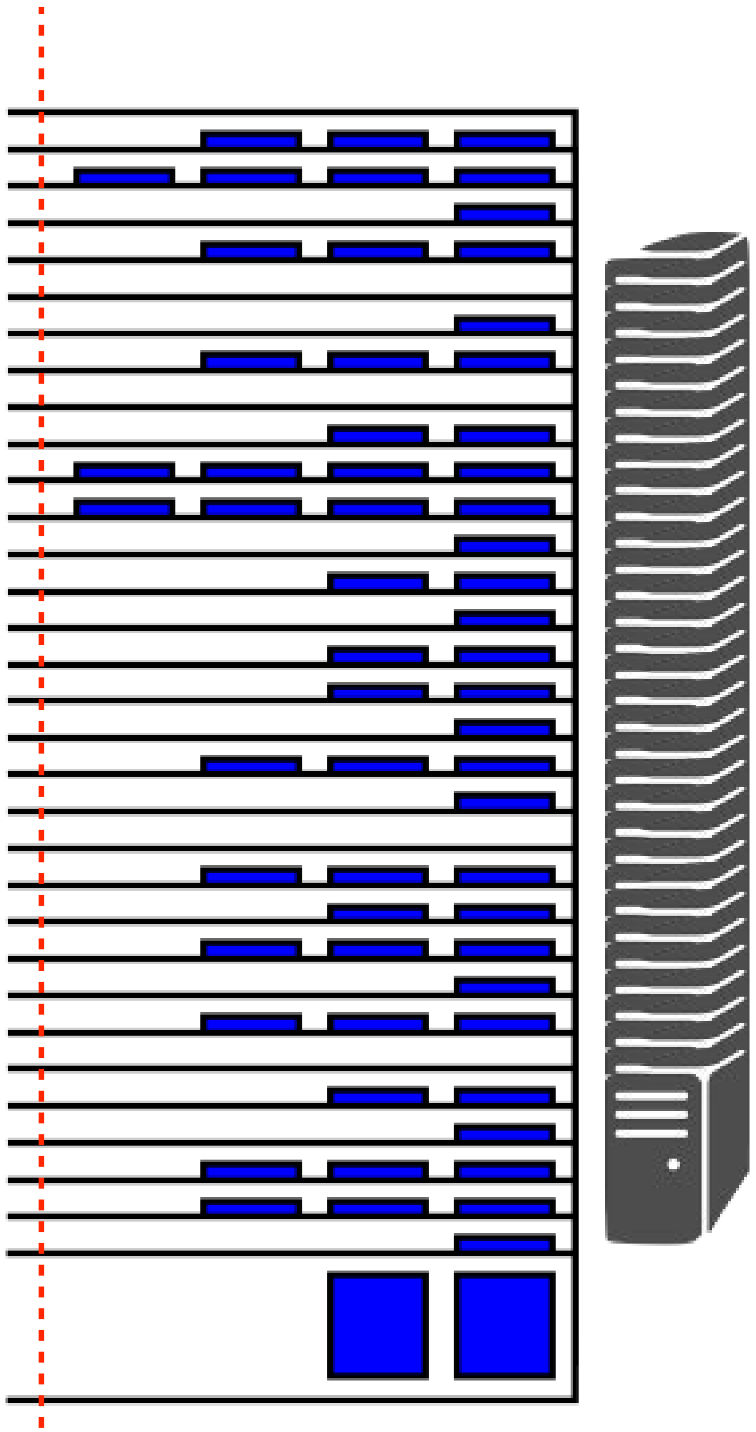
- Many-server heavy-traffic delay

$$\boxed{\frac{\log \frac{1}{1-\rho}}{\log(1+d)}}$$

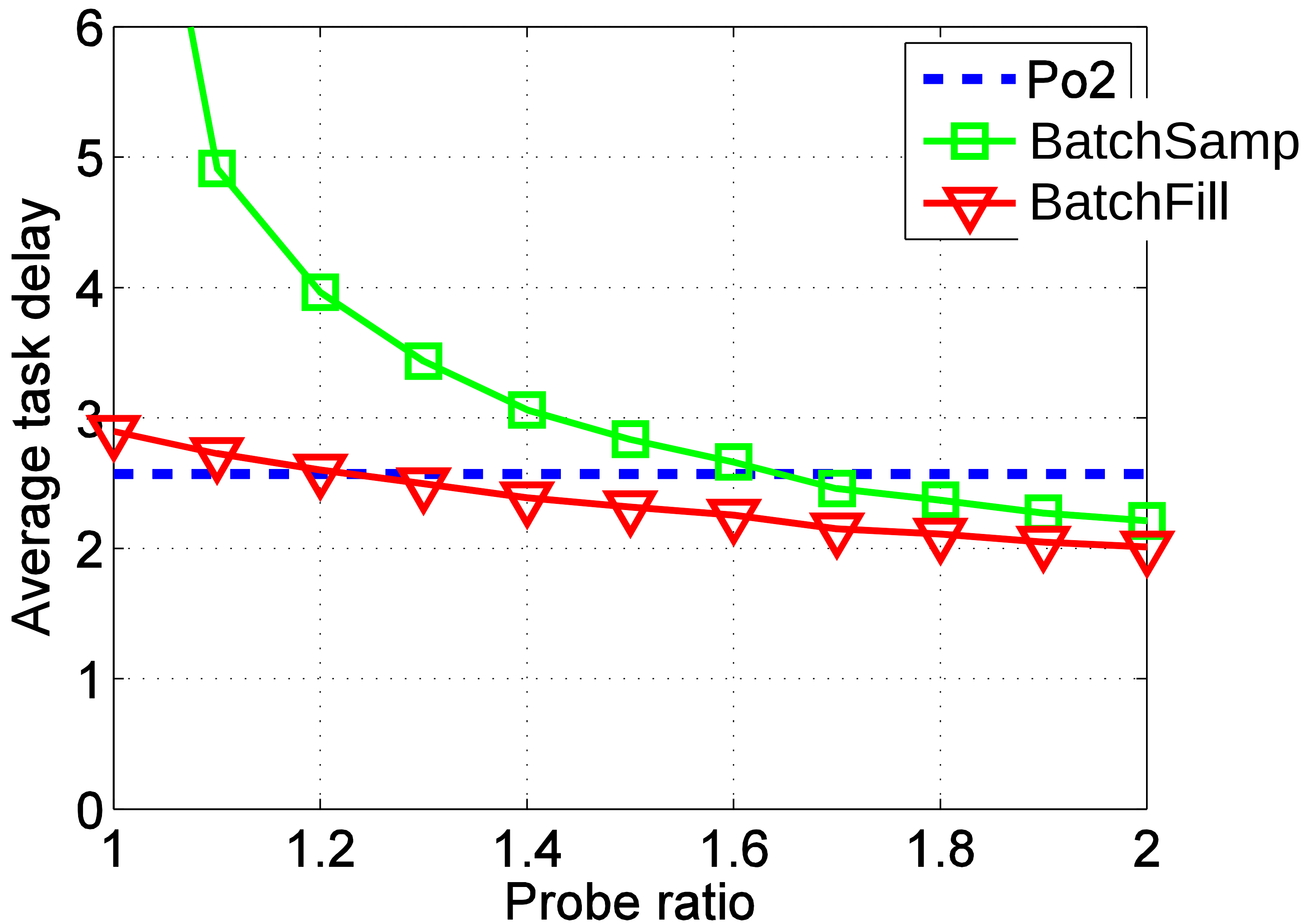
- Queue length upper-bounded by

$$\boxed{\left\lceil \frac{\log \frac{1}{1-\rho}}{\log(1+\rho d)} \right\rceil}$$

$\rho = 0.9$, $d = 1.1$, upper bound 4

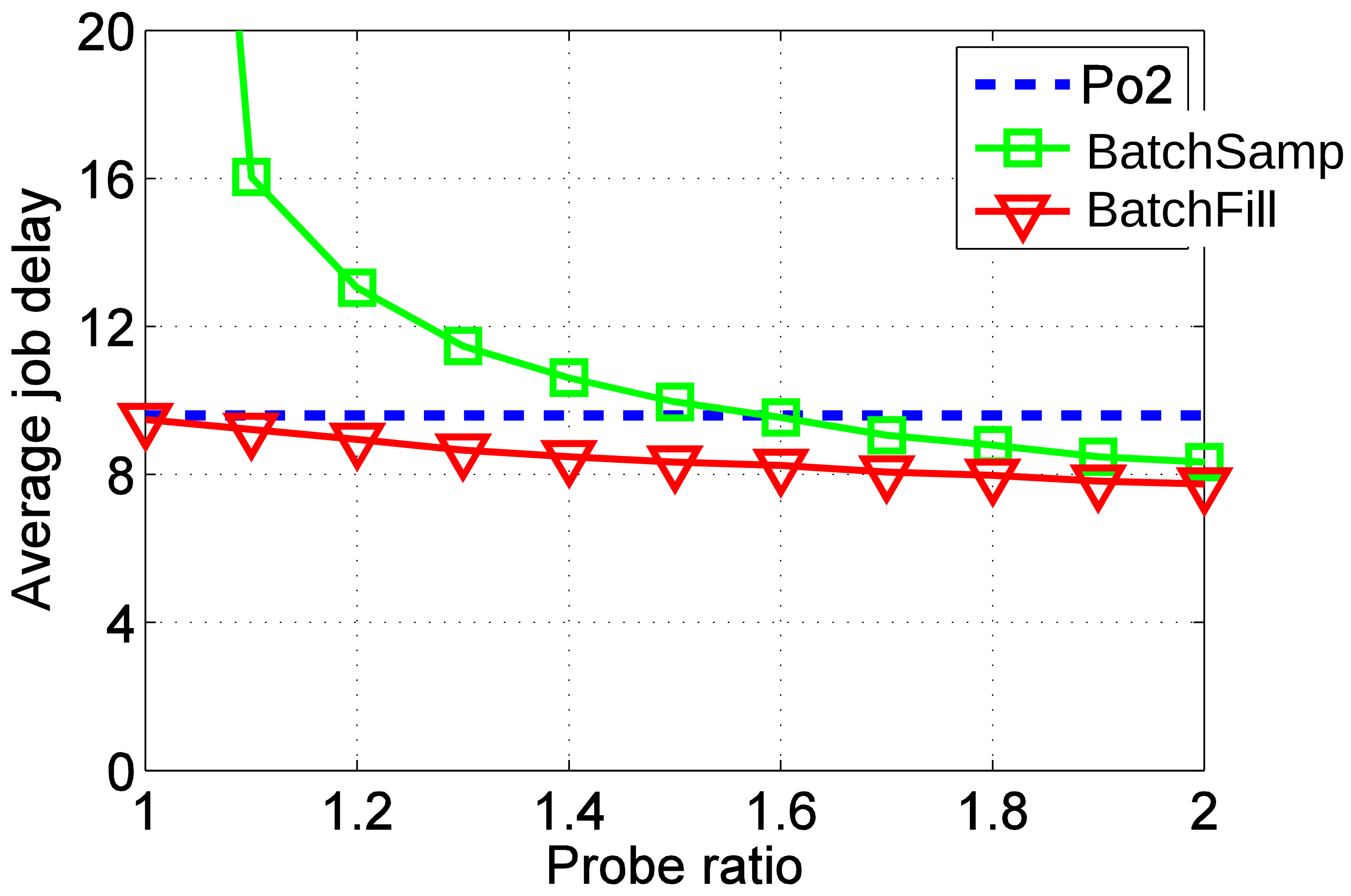


Simulation Results: Average Task Delay



n = 10,000, m = 100, ρ = 0.7

Simulation Results: Average Job Delay



n = 10,000, m = 100, ρ = 0.7

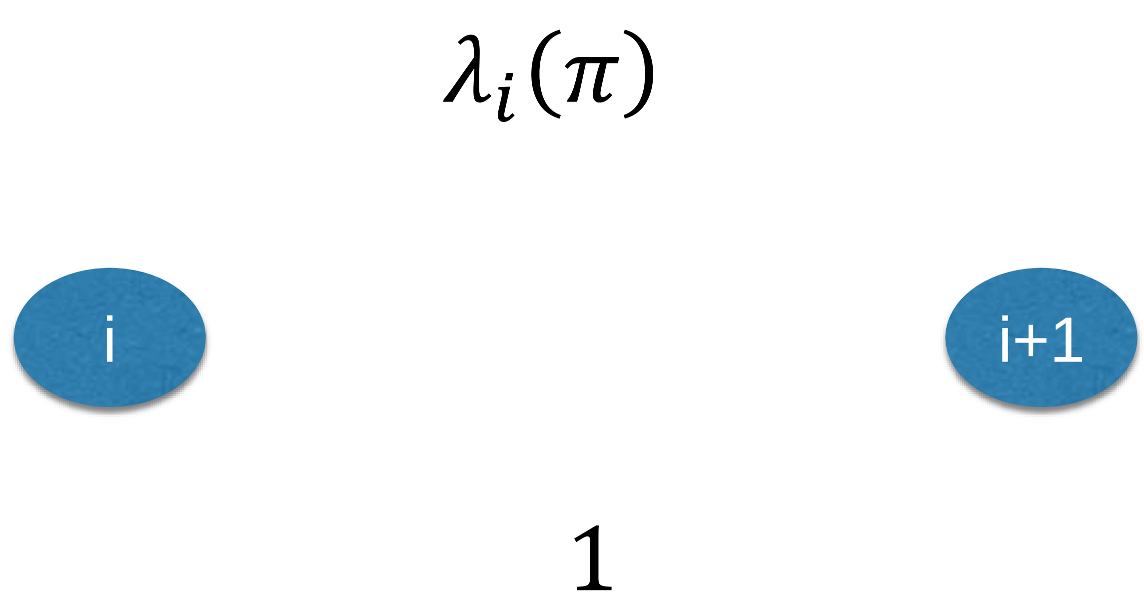
Mean Field Approximation

General Idea

- Focus on a particular queue: Assume all other queues are in steady-state (stationary distribution $\boldsymbol{\pi}$) and independent of each other

- The arrival rate when there are i tasks in the queue is a function of the state of the other queues, and hence a function of $\boldsymbol{\pi}$

- $\boldsymbol{\lambda_i(\pi)\pi_i = \pi_{i+1}}$



Calculating $\lambda_i(\pi)$: Po2

- Task arrival rate $n\rho$
- Prob(a particular queue is sampled) = $2/n$
- Probability being chosen $\pi_i/2+\pi_{i+1}+\pi_{i+2}+\ldots$
- $\lambda_i = \rho(\pi_i + 2\pi_{i+1} + 2\pi_{i+2} + \cdots)$

Stationary Distribution: Po2

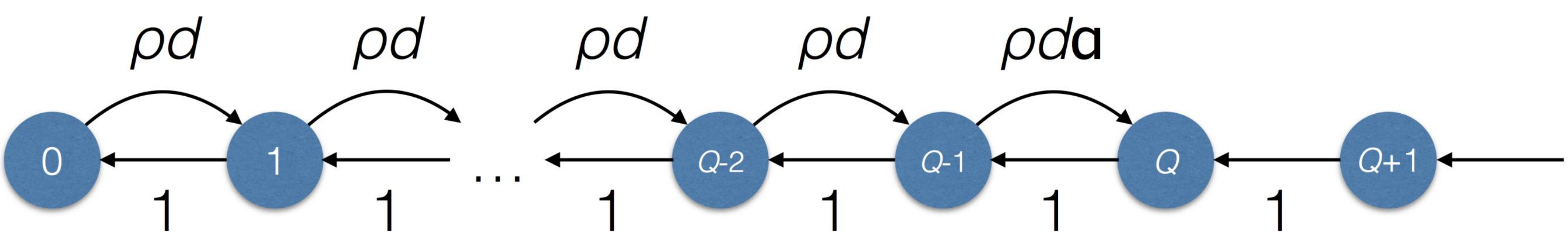
- Queue length distribution: $\boxed{\pi_i = \rho^{2^i - 1} - \rho^{2^{i+1} - 1}}$

- Delay: $\dfrac{\sum_{i=1}^{\infty} i\pi_i}{\rho} = \boxed{\sum_{i=1}^{\infty} \rho^{2^i - 2}}$

BatchSamp

- No waterfilling
- Reducible to finite states

$0 < a \leq 1$



Why is the queue finite?

- md out of n queues are sampled
- We see the empirical distribution: $md\pi_0$ empty queues, $md\pi_1$ queues with one task, etc.

$$0, 0, 0, 1, 1, 1, 1, 1, \dots, j, j, j, j, \dots$$

- The first m of these queues gets one task each
- Under the mean-field approximation, the number of tasks in the m^{th} queue is fixed, as a function of π

π for BatchSamp

- $\pi_0 = 1-\rho$
- $\pi_i = (1-\rho)\rho^i d^i, \quad 1 \le i \le Q-1$
- $\pi_Q = 1-(1-\rho)(\rho^Q d^Q-1)/(\rho d-1)$

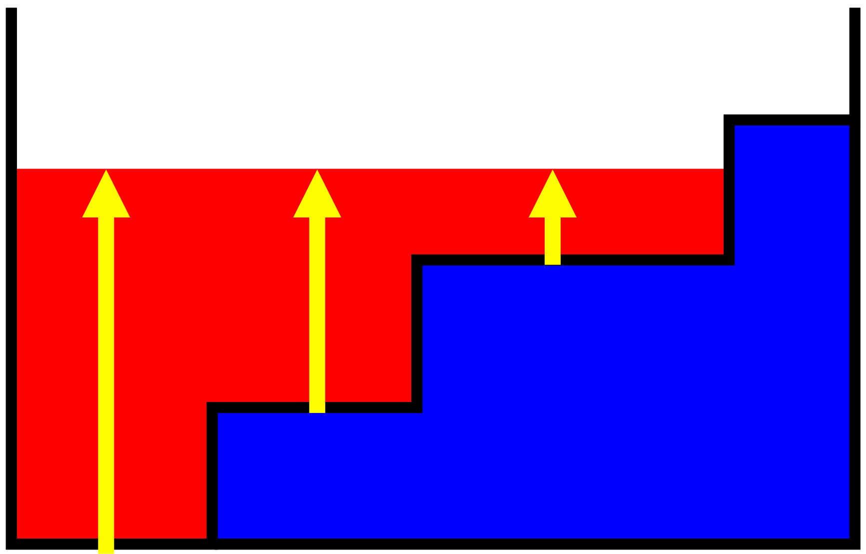
Cutoff queue length

$$\boxed{Q = \left\lceil \frac{\log \frac{d-1}{d(1-\rho)}}{\log(\rho d)} \right\rceil}$$

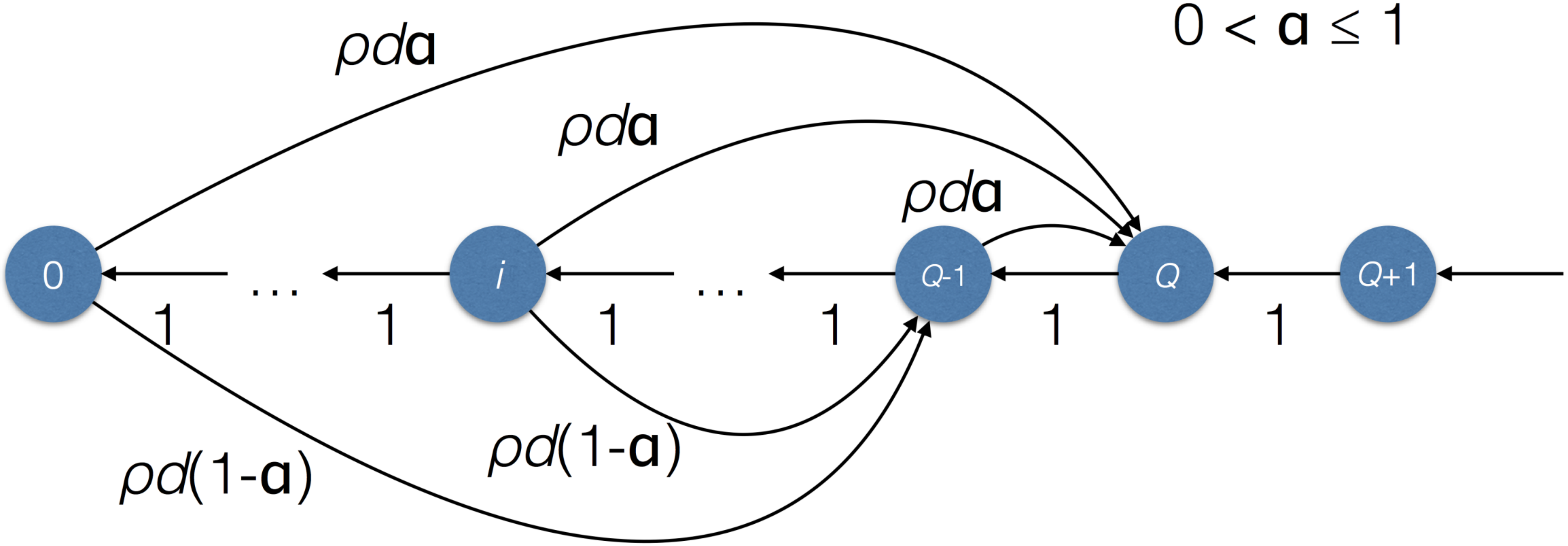
- Delay $\dfrac{\sum_{i=1}^{\infty} i\pi_i}{\rho} \approx \boxed{\dfrac{\log \frac{1}{1-\rho}}{\log d}}$ (heavy-traffic)
- When d = 2, asymptotically equivalent to Po2

BatchFill

- Constant up-crossing rate to Q or Q-1, determined by the stationary distribution $\boldsymbol{\pi}$



$0 < \mathbf{a} \leq 1$



π for BatchFill

- $\pi_0 = 1-\rho$
- $\pi_i = (1-\rho)\rho d(1+\rho d)^{i-1}, \quad 1 \le i \le Q-1$
- $\pi_Q = 1-(1-\rho)(1+\rho d)^{Q-1}$

Cutoff queue length

$$\boxed{Q = \left\lceil \frac{\log \frac{1}{1-\rho}}{\log(1+\rho d)} \right\rceil}$$

- Delay $$\frac{\sum_{i=1}^{\infty} i\pi_i}{\rho} \approx \boxed{\frac{\log \frac{1}{1-\rho}}{\log(1+d)}}$$ (heavy-traffic)

- Better asymptotic delay than Po2 when $d > 1$!

Justifying the Mean-Field Approximation

Ordinary Differential Equation (ODE) for BatchFill

- Deriving an ODE: Derivative is given by

$$\lim_{n\to\infty} \lim_{\delta\to 0^+} E\left(\frac{change\ in\ state}{\delta} \middle| current\ state\right)$$

- ODE

$$\frac{dx_i}{dt} = \begin{cases} -(1+\rho d)x_i + x_{i+1} & i \le \bar{X}_x - 2 \\ \rho d(1-\alpha_x)\sum_{j=0}^{i-1} x_j - (1+\rho d\alpha_x)x_i + x_{i+1} & i = \bar{X}_x - 1 \\ \rho d\alpha_x \sum_{j=0}^{i} x_j - x_i + x_{i+1} & i = \bar{X}_x \\ -x_i + x_{i+1} & \text{otherwise} \end{cases}$$

Global Asymptotic Stability of ODE

- $s_i(t)$ fraction of servers with queue size $\geq i$
- Lyapunov function

$$V(s) = \sum_{i=1}^{\infty} |s_i - \hat{s}_i|$$

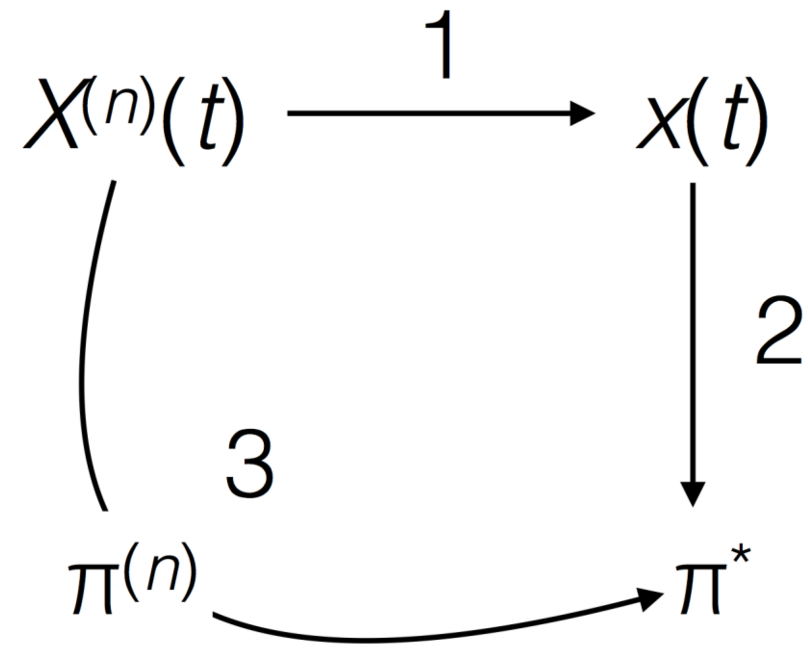
- $s(t)$ converges to the equilibrium point for any $s(0)$

Justifying the Mean-Field Approximation

1. The ODE approximation (in a finite time interval) works well. The deviation from the ODE goes to zero as *n* goes to infinity

2. Global asymptotic stability of the ODE

3. Interchange of limits



Conclusions

- One sample can be powerful in randomized load-balancing
- Batch arrivals can be exploited to reduce sampling overhead
- Extensions: (i) Variable batch sizes, (ii) Batch arrivals are not necessary, (iii) General service-time distributions and Processor Sharing