Electronic Companion: Flexible Queueing Architectures[9]

John N. Tsitsiklis

LIDS, Massachusetts Institute of Technology, Cambridge, MA 02139, jnt@mit.edu

Kuang Xu

Graduate School of Business, Stanford University, Stanford, CA 94305, kuangxu@stanford.edu

Appendix A: Proofs

A.1. Proof of Lemma 3.2

Proof. Fix $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n) \in \boldsymbol{\Lambda}_n(u_n)$, and let g_n be a $(\gamma/\beta_n, \beta_n)$-expander, where $\gamma > \rho$ and $\beta_n \geq u_n$. By the max-flow-min-cut theorem, and the fact that all servers have unit capacity, it suffices to show that

$$\sum_{i \in S} \lambda_i < |\mathcal{N}(S)|, \quad \forall S \subset I. \tag{30}$$

We consider two cases, depending on the size of S.

1. Suppose that $|S| < \gamma n/\beta_n$. By the expansion property of g_n, we have that

$$\mathcal{N}(S) \geq \beta_n |S| \geq u_n |S| > \sum_{i \in S} \lambda_i, \tag{31}$$

   where the second inequality follows from the fact that $\beta_n \geq u_n$, and the last inequality from $\lambda_i < u_n$ for all $i \in I$.

2. Suppose that $|S| \geq \gamma n/\beta_n$. By removing, if necessary, some of the nodes in S, we obtain a set $S' \subset S$ of size exactly $\gamma n/\beta_n$, and

$$\mathcal{N}(S) \geq \mathcal{N}(S') \overset{(a)}{\geq} \gamma n > \rho n \overset{(b)}{\geq} \sum_{i \in S} \lambda_i, \tag{32}$$

   where step (a) follows from the expansion property, and step (b) from the assumption that $\sum_{i \in I} \lambda_i \leq \rho n$.

This completes the proof. Q.E.D.

A.2. Proof of Lemma 3.3

Proof. Lemma 3.3 is a consequence of the following standard result (cf. [1]), where we let $d = d_n$, $\beta = \beta_n$, and $\alpha = \gamma/\beta_n = \sqrt{\rho}/\beta_n$, and observe that $\log_2 \beta_n \ll \beta_n$ as $n \to \infty$.

[9] May 2015; revised October 2016.

Lemma A.1 *Fix $n \geq 1$, $\beta \geq 1$ and $\alpha\beta < 1$. If*

$$d \geq \frac{1 + \log_2 \beta + (\beta + 1)\log_2 e}{-\log_2(\alpha\beta)} + \beta + 1, \tag{33}$$

then there exists an (α, β)-expander with maximum degree d.

Q.E.D.

A.3. Proof of Theorem 3.5

Proof. Since the arrival rate vector $\boldsymbol{\lambda}_n$ whose existence we want to show can depend on the architecture, we assume, without loss of generality, that servers and queues are clustered in the same manner: server i and queue i belong to the same cluster. Since all servers have capacity 1, and each cluster has exactly d_n servers, it suffices to show that there exists $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n) \in \boldsymbol{\Lambda}_n(u_n)$, such that the total arrival rate to the first queue cluster exceeds d_n, i.e.,

$$\sum_{i=1}^{d_n} \lambda_i > d_n. \tag{34}$$

To this end, consider the vector $\boldsymbol{\lambda}$ where $\lambda_i = \min\{2, (1+u_n)/2\}$ for all $i \in \{1, \ldots, d_n\}$, and $\lambda_i = 0$ for $i \geq d_n + 1$. Because of the assumption $u_n > 1$ in the statement of the theorem, we have that

$$\max_{1 \leq i \leq n} \lambda_i = \min\{2, (1+u_n)/2\} \leq \frac{1+u_n}{2} < u_n, \tag{35}$$

and

$$\sum_{i=1}^{n} \lambda_i = d_n \min\{2, (1+u_n)/2\} \leq 2d_n \leq 2 \cdot \frac{\rho}{2} n = \rho n, \tag{36}$$

where the last inequality in Eq. (36) follows from the assumption that $d_n \leq \frac{\rho}{2} n$. Eqs. (35) and (36) together ensure that $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}_n(u_n)$ (cf. Condition 1). Since we have assumed that $u_n > 1$, we have $\lambda_i > 1$, for $i = 1, \ldots, d_n$, and therefore Eq. (34) holds for this $\boldsymbol{\lambda}$. We thus have that $\boldsymbol{\lambda} \notin \mathbf{R}(g_n)$, which proves our claim. Q.E.D.

A.4. Proof of Theorem 3.6

Proof. **Part (a); Eq.** (5). We will use the following classical result due to Hoeffding, adapted from Theorem 3 in [4].

Lemma A.2 *Fix integers m and n, where $0 < m < n$. Let $X_1, X_2, \ldots, X_m$ be random variables drawn uniformly from a finite set $C = \{c_1, \ldots, c_n\}$, without replacement. Suppose that $0 \leq c_i \leq b$ for all i, and let $\sigma^2 = \mathrm{Var}(X_1)$. Let $\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$. Then,*

$$\mathbb{P}\left(\overline{X} \geq \mathbb{E}\left(\overline{X}\right) + t\right) \leq \exp\left(-\frac{mt}{b}\left[\left(1 + \frac{\sigma^2}{bt}\right)\ln\left(1 + \frac{bt}{\sigma^2}\right) - 1\right]\right), \tag{37}$$

for all $t \in (0, b)$.

We fix some $\boldsymbol{\lambda}_n \in \boldsymbol{\Lambda}_n(u_n)$. If $u_n < 1$, then $\boldsymbol{\lambda}_n \in \boldsymbol{\Lambda}_n(1)$. It therefore suffices to prove the result for the case where $u_n \geq 1$ and we will henceforth assume that this is the case. Recall that $A_k \subset I$ is the set of d_n queues in the kth queue cluster generated by the partition $\sigma_n = (A_1, \ldots, A_{n/d_n})$. We consider some $\epsilon \in (0, 1/\rho)$, and define the event E_k as

$$E_k = \left\{ \sum_{i \in A_k} \lambda_i > (1+\epsilon)\rho d_n \right\}. \tag{38}$$

Since σ_n is drawn uniformly at random from all possible partitions, it is not difficult to see that $\sum_{i \in A_k} \lambda_i$ has the same distribution as $\sum_{i=1}^{d_n} X_i$, where $X_1, X_2, \ldots, X_{d_n}$ are d_n random variables drawn uniformly at random, without replacement, from the set $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. Note that $\epsilon\rho < 1 \leq u_n$, so that $\epsilon\rho \in (0, u_n)$. We can therefore apply Lemma A.2, with $m = d_n$, $b = u_n$, and $t = \epsilon\rho$, to obtain

$$\begin{aligned} \mathbb{P}(E_1) =& \mathbb{P}\left(\sum_{i=1}^{d_n} X_i > (1+\epsilon)\rho d_n \right) \\ \stackrel{(a)}{\leq}& \mathbb{P}\left(\frac{1}{d_n} \sum_{i=1}^{d_n} X_i > \mathbb{E}\left(\frac{1}{d_n} \sum_{i=1}^{d_n} X_i \right) + \epsilon\rho \right) \\ \leq& \exp\left(-\frac{\epsilon\rho d_n}{u_n} \left[\left(1 + \frac{\mathrm{Var}(X_1)}{\epsilon\rho u_n} \right) \ln\left(1 + \frac{\epsilon\rho u_n}{\mathrm{Var}(X_1)} \right) - 1 \right] \right), \end{aligned} \tag{39}$$

where the probability is taken with respect to the randomness in G, and where in step (a) we used the fact that

$$\mathbb{E}\left(\sum_{i=1}^{d_n} X_i \right) = \sum_{i=1}^{d_n} \mathbb{E}(X_i) = d_n \mathbb{E}(X_1) = d_n \left(\frac{1}{n} \sum_{i=1}^{n} \lambda_i \right) \leq \rho d_n. \tag{40}$$

We now develop an upper bound on $\mathrm{Var}(X_1)$. Since X_1 takes values in $[0, u_n]$, we have $X_1^2 \leq u_n X_1$ and, therefore,

$$\mathrm{Var}(X_1) \leq \mathbb{E}(X_1^2) \leq u_n \mathbb{E}(X_1) \leq \rho u_n. \tag{41}$$

Observe that for all $a, x > 0$,

$$\frac{d}{dx}(1 + x/a)\ln(1 + a/x) = -\frac{1}{x} + \frac{1}{a}\ln(1 + a/x) < -\frac{1}{x} + \frac{1}{a} \cdot \frac{a}{x} = 0. \tag{42}$$

Therefore, with the substitutions $a = \epsilon\rho u_n$ and $x = \mathrm{Var}(X_1)$, we have that the right-hand-side of (39) is increasing in $\mathrm{Var}(X_1)$. Combining Eqs. (39) and (41), we obtain

$$\mathbb{P}(E_1) \leq \exp\left(-\frac{\epsilon\rho d_n}{u_n} \left[\left(1 + \frac{1}{\epsilon} \right) \ln(1+\epsilon) - 1 \right] \right).$$

Note that

$$\frac{d}{dx}\left(1 + \frac{1}{x} \right) \ln(1+x) = \frac{1}{x^2}(x - \ln(1+x)) \stackrel{(a)}{\to} \frac{1}{2}, \quad \text{as } x \downarrow 0, \tag{43}$$

where step (a) follows from applying l'Hôpital's rule. We thus have that $\left[\left(1+\frac{1}{\epsilon}\right)\ln(1+\epsilon)-1\right] \sim \frac{1}{2}\epsilon \geq \frac{1}{3}\epsilon$, as $\epsilon \downarrow 0$, it follows that there exists $\theta > 0$ such that for all $\epsilon \in (0,\theta)$,

$$\mathbb{P}(E_1) \leq \exp\left(-\frac{\rho}{3}\cdot\frac{\epsilon^2 d_n}{u_n}\right). \tag{44}$$

Let $\epsilon = \frac{1}{2}\min\{\frac{1}{\rho}-1,\theta\}$; in particular, our earlier assumption that $\epsilon\rho < 1$ is satisfied. Suppose that $u_n \leq \frac{\rho\epsilon^2}{6} d_n \ln^{-1} n$. Combining Eq. (44) with the union bound, we have that

$$\begin{aligned}\mathbb{P}_{G_n}(\boldsymbol{\lambda}_n \notin \mathbf{R}(G_n)) &\leq \mathbb{P}\left(\bigcup_{k=1}^{n/d_n} E_k\right)\\ &\leq \sum_{k=1}^{n/d_n} \mathbb{P}(E_k)\\ &\leq \frac{n}{d_n}\exp\left(-\frac{\rho}{3}\cdot\frac{\epsilon^2 d_n}{u_n}\right)\\ &\overset{(a)}{\leq} \frac{n}{d_n}\cdot\frac{1}{n^2}\\ &\leq n^{-1},\end{aligned} \tag{45}$$

where step (a) follows from the assumption that $u_n \leq \frac{\rho\epsilon^2}{6} d_n \ln^{-1} n$. It follows that

$$\lim_{n\to\infty} \inf_{\boldsymbol{\lambda}_n \in \boldsymbol{\Lambda}_n(u_n)} \mathbb{P}_{G_n}(\boldsymbol{\lambda}_n \in \mathbf{R}(G_n)) \geq \lim_{n\to\infty}\left(1-\frac{1}{n}\right) = 1. \tag{46}$$

We have therefore proved part (a) of the theorem, with $c_2 = \rho\epsilon^2/6$.

Part (b); Eq. (6).

Let us fix a large enough constant c_3, whose value will be specified later, and let

$$v_n = c_3 \frac{d_n}{\ln n}. \tag{47}$$

For this part of the proof, we will assume that $u_n > v_n$. Because we are interested in showing a result for the worst case over all $\boldsymbol{\lambda}_n \in \boldsymbol{\Lambda}_n(u_n)$, we can assume that $u_n \ll n$.

At this point, we could analyze the model for a worst-case choice of $\boldsymbol{\lambda}_n$. However, the analysis turns out to be simpler if we employ the probabilistic method. Denote by μ_n a probability measure over $\boldsymbol{\Lambda}_n(u_n)$. Let $\boldsymbol{\lambda}_n$ be a random vector drawn from the distribution μ_n, independent of the randomness in the Random Modular architecture, G. (For convenience, we suppress the subscript n and write G instead of G_n.) The following elementary fact captures the essence of the probabilistic method.

Lemma A.3 *Fix n, a measure μ_n on $\boldsymbol{\Lambda}_n(u_n)$, and a constant a_n. Suppose that*

$$\mathbb{P}_{\boldsymbol{\lambda}_n,G}(\boldsymbol{\lambda}_n \notin \mathbf{R}(G)) \geq a_n, \tag{48}$$

where $\mathbb{P}_{\boldsymbol{\lambda}_n,G}$ stands for the product of the measures μ_n (for $\boldsymbol{\lambda}_n$) and $\mathbb{P}_G$ (for G). Then,

$$\sup_{\tilde{\boldsymbol{\lambda}}_n \in \boldsymbol{\Lambda}_n(u_n)} \mathbb{P}_G(\tilde{\boldsymbol{\lambda}}_n \notin \mathbf{R}(G)) \geq a_n. \tag{49}$$

Proof. We have that

$$\begin{aligned} \sup_{\tilde{\boldsymbol{\lambda}}_n \in \boldsymbol{\Lambda}_n(u_n)} \mathbb{P}_G(\tilde{\boldsymbol{\lambda}}_n \notin \mathbf{R}(G)) \geq & \int_{\tilde{\boldsymbol{\lambda}}_n \in \boldsymbol{\Lambda}_n(u_n)} \mathbb{P}_G(\tilde{\boldsymbol{\lambda}}_n \notin \mathbf{R}(G))\, d\mu_n(\tilde{\boldsymbol{\lambda}}_n) \\ = & \mathbb{P}_{\boldsymbol{\lambda}_n,G}\left(\boldsymbol{\lambda}_n \notin \mathbf{R}(G)\right) \\ \geq & a_n. \end{aligned} \tag{50}$$

Q.E.D.

We will now construct sequences, $\{\mu_n : n \in \mathbb{N}\}$, and $\{a_n : n \in \mathbb{N}\}$, with $\lim_{n\to\infty} a_n = 1$, so that Eq. (48) holds for all n. To simplify notation, in the rest of this proof we will write $\mathbb{P}$ instead of $\mathbb{P}_G$ or $\mathbb{P}_{\boldsymbol{\lambda}_n,G}$, etc. Which particular measure we are dealing with will always be clear from the context.

Fix $n \in \mathbb{N}$. We first construct the distribution μ_n. Let $\boldsymbol{\lambda}' = (\lambda'_1, \lambda'_2, \ldots, \lambda'_n)$ be a random vector with independent components and with

$$\lambda'_i = \begin{cases} v_n, & \text{w.p. } \frac{\rho}{(1+\epsilon)v_n}, \\ 0, & \text{otherwise,} \end{cases} \tag{51}$$

for all i. Let H be the event defined by

$$H = \left\{ \sum_{i=1}^{n} \lambda'_i \leq \rho n \right\}. \tag{52}$$

Let $\boldsymbol{\lambda}_n$ be the random vector given by

$$\boldsymbol{\lambda}_n = \mathbb{I}(H)\boldsymbol{\lambda}', \tag{53}$$

where $\mathbf{0}$ is the zero vector of dimension n, and where $\mathbb{I}(\cdot)$ is the indicator function. That is, $\boldsymbol{\lambda}_n$ takes on the value of $\boldsymbol{\lambda}'$ if H occurs, and is set to zero, otherwise. It is not difficult to verify that, by construction, we always have $\boldsymbol{\lambda}_n \in \boldsymbol{\Lambda}_n(u_n)$. We let μ_n be the distribution of this random vector $\boldsymbol{\lambda}_n$.

We next show that

$$\lim_{n\to\infty} \mathbb{P}\left(\boldsymbol{\lambda}_n \notin \mathbf{R}(G)\right) = 1, \tag{54}$$

which, together with Lemma A.3 above, will complete the proof of the theorem. Fix some $\epsilon > \frac{1}{\rho} - 1$, so that $(1+\epsilon)\rho > 1$, and define the event

$$E_k = \left\{ \sum_{i \in A_k} \lambda'_i > (1+\epsilon)\rho d_n \right\}, \qquad k \in \{1, \ldots, n/d_n\}. \tag{55}$$

Note that, if some E_k occurs, then $\boldsymbol{\lambda}'$ will not be in $\mathbb{R}(G)$. Therefore,

$$\mathbb{P}(\boldsymbol{\lambda}' \notin \mathbf{R}(G)) \geq \mathbb{P}\left(\bigcup_{k=1}^{n/d_n} E_k\right). \tag{56}$$

Let $X_1, X_2, \ldots$ be i.i.d. Bernoulli random variables with

$$\mathbb{E}(X_1) = \mathbb{P}(X_1 = 1) = \frac{\rho}{(1+\epsilon)v_n}. \tag{57}$$

By the definition of $\boldsymbol{\lambda}'$ (cf. Eq. (51)), we have that

$$\begin{aligned}\mathbb{P}(E_1) =& \mathbb{P}\left(\sum_{i\in A_1} \lambda_i' > (1+\epsilon)\rho d_n\right)\\ =& \mathbb{P}\left(\sum_{i=1}^{d_n} X_i > (1+\epsilon)\rho \frac{d_n}{v_n}\right)\\ =& \mathbb{P}\left(\frac{1}{d_n}\sum_{i=1}^{d_n} X_i > (1+\epsilon)^2 \mathbb{E}\left(X_1\right)\right).\end{aligned} \tag{58}$$

By Sanov's theorem (cf. Chapter 12 of [2]), we have that

$$\begin{aligned}\mathbb{P}(E_1) =& \mathbb{P}\left(\frac{1}{d_n}\sum_{i=1}^{d_n} X_i > (1+\epsilon)^2 \mathbb{E}\left(X_1\right)\right)\\ \gtrsim& \frac{1}{d_n^2}\exp\left(-D_B\left(\frac{(1+\epsilon)\rho}{v_n}\middle\| \frac{\rho}{(1+\epsilon)v_n}\right) d_n\right),\end{aligned} \tag{59}$$

where $D_B(p\|q)$ is the Kullback-Leibler divergence between two Bernoulli distributions with parameters p and q, respectively:

$$D_B(p\|q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}. \tag{60}$$

Let us fix some $r \in (0,1)$. Using the fact that $\ln(1+y) \sim y$ as $y \to 0$, we have that

$$D_B\left(x\| rx\right) \sim x\left[\ln\frac{1}{r} + (1-r)\right], \quad \text{as } x \to 0. \tag{61}$$

Recall that $d_n \geq c_1 \ln n$ and $v_n \geq / \ln n$. By Eq. (61), with $x = (1+\epsilon)\rho/v_n$, $r = 1/(1+\epsilon)^2$, and for the given c_1, we can set c_3 to be sufficiently large so that

$$\begin{aligned}D_B\left(\frac{(1+\epsilon)\rho}{v_n}\middle\| \frac{\rho}{(1+\epsilon)v_n}\right) \leq& 2\frac{(1+\epsilon)\rho}{v_n}\cdot\left[\ln(1+\epsilon)^2 + \left(1 - \frac{1}{(1+\epsilon)^2}\right)\right]\\ =& \frac{2h}{v_n},\end{aligned} \tag{62}$$

for all sufficiently large n, where $h = (1+\epsilon)\rho\left[\ln(1+\epsilon)^2 + \left(1-\frac{1}{(1+\epsilon)^2}\right)\right] > 0$. Combining Eqs. (59) and (62), we have that

$$\mathbb{P}(E_1) \gtrsim \frac{1}{d_n^2}\exp\left(-2h\frac{d_n}{v_n}\right) \overset{(a)}{\gtrsim} \frac{1}{d_n^2} n^{-2h/c_3}, \tag{63}$$

where step (a) follows from the assumption that $v_n \geq c_3 d_n / \ln n$. Equation (63) can be rewritten in the form

$$\mathbb{P}(E_1) \geq \frac{c}{d_n^2} n^{-2h/c_3}, \tag{64}$$

where c is a positive constant, and where the inequality is valid for large enough n.

Fix $c_3 = 40h$, and recall that $\epsilon > \frac{1}{\rho} - 1$. We have that

$$\begin{aligned}
\mathbb{P}(\boldsymbol{\lambda}' \notin \mathbf{R}(G)) \geq & \mathbb{P}\left(\bigcup_{k=1}^{n/d_n} E_k\right) \\
\overset{(a)}{=} & 1 - \prod_{k=1}^{n/d_n} \left(1 - \mathbb{P}(E_k)\right) \\
= & 1 - \left(1 - \mathbb{P}(E_1)\right)^{n/d_n} \\
\overset{(b)}{\geq} & 1 - \left(1 - c d_n^{-3} n^{1-2h/c_3} d_n / n\right)^{n/d_n} \\
\overset{(c)}{\geq} & 1 - \left(1 - c n^{0.05} d_n / n\right)^{n/d_n} \\
\to & 1, \quad \text{as } n \to \infty,
\end{aligned} \tag{65}$$

where step (a) is based on the independence among the events E_k, which is in turn based on the independence among the λ_i's; step (b) follows from Eq. (64) and some rearrangement; step (c) follows from the assumption in the statement of the theorem that $d_n \leq n^{0.3}$, and our choice of $c_3 = 40h$.

We next show that the event H occurs with high probability when n is large. Let, as before, the X_is be i.i.d. Bernoulli random variables with $\mathbb{E}(X_1) = \frac{\rho}{v_n(1+\epsilon)}$. Then,

$$\begin{aligned}
\mathbb{P}(H) = & \mathbb{P}\left(\sum_{i=1}^{n} \lambda_i' \leq \rho n\right) \\
= & \mathbb{P}\left(\sum_{i=1}^{n} X_i \leq \rho n / v_n\right) \\
= & \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq (1+\epsilon)\mathbb{E}\left(X_1\right)\right) \to 1, \quad \text{as } n \to \infty,
\end{aligned} \tag{66}$$

by the weak law of large numbers.

We are now ready to prove Eq. (54). We have that

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\lambda}_n, G}\Big(\boldsymbol{\lambda}_n \notin \mathbf{R}(G)\Big) = & \mathbb{P}_{\boldsymbol{\lambda}', G}\Big(\mathbb{I}(H)\boldsymbol{\lambda}' \notin \mathbf{R}(G)\Big) \\
= & \mathbb{P}_{\boldsymbol{\lambda}', G}\Big(H \cap \{\boldsymbol{\lambda}' \notin \mathbf{R}(G)\}\Big) \\
\geq & \mathbb{P}(H) + \mathbb{P}\big(\boldsymbol{\lambda}' \notin \mathbf{R}(G)\big) - 1 \\
\to & 1, \quad \text{as } n \to \infty,
\end{aligned} \tag{67}$$

where the last step follows from Eqs. (65) and (66). By Lemma A.3, Eq. (67) implies that $\lim_{n\to\infty}\sup_{\boldsymbol{\lambda}_n\in\boldsymbol{\Lambda}_n(u_n)}\mathbb{P}_{G_n}(\boldsymbol{\lambda}_n\notin\mathbf{R}(G)) = 1$, which is in turn equivalent to $\lim_{n\to\infty}\inf_{\boldsymbol{\lambda}_n\in\boldsymbol{\Lambda}_n(u_n)}\mathbb{P}_{G_n}(\boldsymbol{\lambda}_n\in\mathbf{R}(G)) = 0$. This proves Eq. (6). Q.E.D.

A.5. Proof of Theorem 3.7

Proof. Denote by $Q_i(t)$ the number of jobs in queue i at time t, and by $Q_k(t)$ the total number of jobs in queue cluster k, i.e.,

$$Q_k(t) = \sum_{i\in A_k} Q_i(t). \tag{68}$$

We note that $Q_k(\cdot)$ is the number of jobs in an $M/M/c$ queue, with $c = d_n$ and arrival rate $\eta_k = \sum_{i\in A_k}\lambda_i$. Also note that since $\boldsymbol{\lambda}_n \in \gamma\mathbf{R}(g_n)$, we have that $\eta_k \le \gamma d_n$. Using the formula for the expected waiting time in queue for an $M/M/c$ queue (cf. Section 2.3 of [3]), one can show that the average waiting time across jobs arriving to cluster k, W_k, satisfies

$$\mathbb{E}\left(W_k|\boldsymbol{\lambda}\right) = \frac{1}{\sum_{i\in A_k}\lambda_i}\sum_{i\in A_k}\lambda_i\mathbb{E}\left(W_i\right) = \frac{C(d_n,\eta_k)}{d_n-\eta_k} \le \frac{C(d_n,\gamma d_n)}{(1-\gamma)d_n} \lesssim \exp(-b\cdot d_n), \tag{69}$$

where $C(c,r)$ is given by

$$C(c,r) = \frac{r^c}{c!}\cdot\frac{1}{c(1-r/c)^2}\left(\frac{r^c}{c!}\cdot\frac{1}{1-r/c}+\sum_{i=0}^{c-1}\frac{r^i}{i!}\right)^{-1}.$$

The last inequality in Eq. (69) follows from the fact that for any given $\gamma\in(0,1)$, there exists $b>0$, so that $C(x,\gamma x)\lesssim \exp(-b\cdot x)$ as $x\to\infty$, as can be checked through elementary algebraic manipulations. Q.E.D.

A.6. Lower Bound on the Total Arrival Rate

We show in this section that the assumption that $\rho\in(1/2,1)$ and $\sum_{i=1}^n\lambda_i \ge (1-\rho)n$ (cf. Eq. (10) in Assumption 4.1) can be made without loss of generality. Fix the traffic intensity $\rho\in(0,1)$, and suppose that $\boldsymbol{\lambda}\in\boldsymbol{\Lambda}_n(u_n)$. Define

$$\rho' = \rho + \frac{1}{2}(1-\rho) = \frac{1+\rho}{2}. \tag{70}$$

Note that $1/2<\rho'<1$, and $1-\rho' = (1-\rho)/2$. Consider a modified vector $\boldsymbol{\lambda}'$, where $\lambda'_i = (1-\rho')+\lambda_i$, for all $i\in\{1,\ldots,n\}$. By construction, we have that

$$\sum_{i=1}^n \lambda'_i \ge (1-\rho')n, \tag{71}$$

$$\sum_{i=1}^n \lambda'_i \le (1-\rho')n + \sum_{i=1}^n\lambda_i \le (1-\rho')n+\rho n = \rho' n, \tag{72}$$

$$\max_{1\le i\le n}\lambda'_i \le \max_{1\le i\le n}\lambda_i + (1-\rho') < u_n + (1-\rho'). \tag{73}$$

The above definition of $\boldsymbol{\lambda}'$ amounts to the following: we feed each queue with an additional independent Poisson stream of artificial (dummy) jobs of rate $1-\rho'$. By Eqs. (72) and (73), the resulting arrival rate vector, $\boldsymbol{\lambda}'$, will belong to the set $\boldsymbol{\Lambda}_n(u_n + 1 - \rho')$. Also, by Eq. (71), it will satisfy the lower bound (10) on the total arrival rate, albeit with a modified traffic intensity of $\rho' \in (1/2, 1)$. Therefore, our assumption can always be satisfied by the insertion of dummy jobs. Note that the increment of $1-\rho'$ to the value of u_n is insignificant in our regime of interest, where $u_n \gg 1$, and the insertion of dummy jobs only requires knowledge of the original traffic intensity, ρ.

A.7. Proof of Lemma 4.5

Proof. Note that because there are ρb_n jobs in a batch, the size of Γ is at most ρb_n, which is in turn less than m_n. This guarantees that the cardinality of $\hat{\Gamma}$ can be taken to be m_n. It therefore suffices to show that

$$\mathbb{P}\left(\max_{1\le i\le n} A_i \ge \hat{u}_n\right) \le 1/n^3. \tag{74}$$

There is a total of ρb_n arriving jobs in a single batch, and for each arriving job

$$\mathbb{P}(\text{the job arrives to queue } i) = \frac{\lambda_i}{\sum_{i=1}^n \lambda_i} \overset{(a)}{\le} \frac{\lambda_i}{(1-\rho)n} \le \frac{u_n}{(1-\rho)n} \overset{(b)}{\le} \frac{1}{2n}\beta_n \le \frac{1}{2n\hat{\rho}}\beta_n, \tag{75}$$

for all i, where steps (a) and (b) follow from the assumptions that $\sum_{i=1}^n \lambda_i \ge (1-\rho)n$ (Eq. (10) in Assumption 4.1) and that $u_n \le \frac{1-\rho}{2}\beta_n$ (in the statement of Theorem 3.4), respectively. From Eq. (75), A_i is stochastically dominated by a binomial random variable $\tilde{A} \overset{d}{=} \text{Bino}(\rho b_n, \frac{1}{2n\hat{\rho}}\beta_n)$, with

$$\mathbb{E}\left(\tilde{A}\right) = \rho b_n \frac{1}{2n\hat{\rho}}\beta_n = \frac{1}{2}\left(\beta_n \frac{\rho b_n/\hat{\rho}}{n}\right) = \frac{1}{2}\left(\beta_n \frac{m_n}{n}\right) = \frac{1}{2}\,\hat{u}_n. \tag{76}$$

Based on this expression of $\mathbb{E}\left(\tilde{A}\right)$, we will now use an exponential tail bound to bound the probability of the event $\{\max_{1\le i\le n} A_i \ge \hat{u}_n\}$. Recall that $b_n = \frac{320}{(1-\rho)^2}\cdot\frac{n\ln n}{\beta_n}$. Using the union bound, we have that

$$\begin{aligned}
\mathbb{P}\left(\max_{1\le i\le n} A_i \ge \hat{u}_n\right) &= \mathbb{P}(A_i \ge \hat{u}_n, \text{ for some } i) \\
&\le n\mathbb{P}(A_1 \ge \hat{u}_n) \\
&\le n\mathbb{P}\left(\tilde{A} \ge \hat{u}_n\right) \\
&\overset{(a)}{=} n\mathbb{P}\left(\tilde{A} \ge 2\mathbb{E}(\tilde{A})\right) \\
&\overset{(b)}{\le} n\exp\left(-\frac{1}{3}\mathbb{E}(\tilde{A})\right) \\
&= n\exp\left(-\frac{\rho}{6\hat{\rho}}\cdot\frac{b_n\beta_n}{n}\right) \\
&\le n\exp\left(-\frac{\rho}{6}\cdot\frac{b_n\beta_n}{n}\right)
\end{aligned} \tag{77}$$

$$
\begin{aligned}
&= n\exp\left(-\frac{\rho}{6}\cdot\frac{320}{(1-\rho)^2}\cdot\frac{n\ln n}{\beta_n}\cdot\frac{\beta_n}{n}\right)\\
&\overset{(c)}{\leq} n\exp\left(-\frac{160}{6}\ln n\right)\\
&\leq n^{-3}.
\end{aligned}
\tag{78}
$$

Step (a) follows from Eq. (76). Step (b) follows from the following multiplicative form of the Chernoff bound (cf. Chapter 4 of [5]), with $\delta=1$: $\mathbb{P}(\tilde{A}\geq(1+\delta)\mu)\leq\exp(-\frac{\delta^2}{2+\delta}\mu)$, where $\tilde{A}$ is a binomial random variable with $\mathbb{E}(\tilde{A})=\mu$. Step (c) follows from the assumption $\rho\in(1/2,1)$ (cf. Assumption 4.1), and hence

$$\frac{\rho}{(1-\rho)^2}\geq\rho\geq 1/2. \tag{79}$$

This completes the proof of Lemma 4.5. Q.E.D.

A.8. Proof of Lemma 4.7

Proof. For a set $S\subset\hat{\Gamma}$, denote by $\mathcal{N}^*(S)$ the set of neighbors of S in $\hat{G}$, i.e., $\mathcal{N}^*(S)=\mathcal{N}(S)\cap\Delta$. To prove Lemma 4.7, we will leverage the fact that the underlying connectivity graph, g_n, is an expander graph with appropriate expansion. As a result, most subsets $S\subset\hat{\Gamma}$ have a large set of neighbors, $\mathcal{N}(S)$, in g_n. Because each server in $\mathcal{N}(S)$ belongs to $\mathcal{N}^*(S)$ independently, as a consequence of our scheduling policy, we will then use a concentration inequality to show that, with high probability, the sizes of the sets $\mathcal{N}^*(S)$ remain sufficiently large. Using the union bound over the relevant sets S, we will finally conclude that $\hat{G}$ has the desired expansion property, with high probability.

By the definition of a $(\gamma/\hat{u}_n,\hat{u}_n)$-expander, we are only interested in the expansion of subsets of $\hat{\Gamma}$ with size less than or equal to $|\hat{\Gamma}|\gamma/\hat{u}_n$. We first verify below that the size of such subsets S is sufficiently small to be able to exploit the expansion property of g_n and to infer that $\mathcal{N}^*(S)$ is large. We have

$$\frac{n\gamma/\beta_n}{|\hat{\Gamma}|\gamma/\hat{u}_n}=\frac{n}{|\hat{\Gamma}|}\cdot\frac{\hat{u}_n}{\beta_n}=\frac{n}{m_n}\cdot\frac{\beta_n\frac{m_n}{n}}{\beta_n}=1, \tag{80}$$

which is equivalent to saying

$$s\leq\gamma n/\beta_n,\quad\forall s\leq|\hat{\Gamma}|\gamma/\hat{u}_n, \tag{81}$$

as desired.

For a set $S\subset\hat{\Gamma}$, we now characterize the size of its neighborhood in $\hat{G}$, $|\mathcal{N}^*(S)|$, which depends on the distribution of the random subset, Δ. Fix some $s\in\mathbb{N}$ with $s\leq|\hat{\Gamma}|\gamma/\hat{u}_n$. From Eq. (81), we know that $s \le \gamma n/\beta_n$. Consider some $S \subset \hat{\Gamma}$ with $|S| = s$. Using the expansion property of g_n, we have that $|\mathcal{N}(S)| \ge \beta_n s$. Therefore,

$$\begin{aligned}
\mathbb{P}(|\mathcal{N}^*(S)| \le \hat{u}_n s) =& \mathbb{P}\left(\sum_{j\in\mathcal{N}(S)} \mathbb{I}(j\in\Delta) \le \hat{u}_n s\right) \\
\overset{(a)}{\le}& \mathbb{P}\left(\text{Bino}\left(|\mathcal{N}(S)|\,, \frac{b_n}{n}(\rho+3\epsilon/4)\right) \le \hat{u}_n s\right) \\
\overset{(b)}{\le}& \mathbb{P}\left(\text{Bino}\left(\beta_n s\,, \frac{b_n}{n}(\rho+3\epsilon/4)\right) \le \hat{u}_n s\right),
\end{aligned} \tag{82}$$

for all sufficiently large n. Step (a) follows from the assumption that $\mathbb{P}(j\in\Delta) \ge (\rho+3\epsilon/4)\frac{b_n}{n}$, and step (b) from the inequality $|\mathcal{N}(S)| \ge \beta_n s$. We observe that

$$\begin{aligned}
\mu \overset{\triangle}{=}& \mathbb{E}\left(\text{Bino}\left(\beta_n s\,, \frac{b_n}{n}(\rho+3\epsilon/4)\right)\right) \\
=& (\rho+3\epsilon/4)\frac{\beta_n b_n}{n} s \\
\overset{(a)}{=}& (\rho+3\epsilon/4)\frac{1}{n}\cdot\frac{80}{\epsilon^2}\cdot\frac{n\ln n}{\beta_n}\beta_n s \\
=& (\rho+3\epsilon/4)\frac{80\ln n}{\epsilon^2} s,
\end{aligned} \tag{83}$$

where in step (a) we used the substitution $b_n = \frac{80}{\epsilon^2}\cdot\frac{n\ln n}{\beta_n}$. We also have that

$$\begin{aligned}
\hat{u}_n =& \beta_n \frac{m_n}{n} \\
=& \beta_n \frac{\rho b_n}{\hat{\rho} n} \\
=& \beta_n \frac{\rho}{\hat{\rho} n}\cdot\frac{80}{\epsilon^2}\cdot\frac{n\ln n}{\beta_n} \\
=& \frac{\rho}{\hat{\rho}}\cdot\frac{80\ln n}{\epsilon^2}.
\end{aligned} \tag{84}$$

By combining Eqs. (83) and (84), we can derive a useful lower bound on the quantity $1-\frac{s\hat{u}_n}{\mu}$, which is recorded in the lemma that follows.

Lemma A.4 *We have that*

$$1-\frac{s\hat{u}_n}{\mu} \ge \frac{\epsilon}{2}. \tag{85}$$

Proof. Using Eqs. (83) and (84) in the first step below, we have that

$$1-\frac{s\hat{u}_n}{\mu} = 1-\frac{\rho}{\hat{\rho}(\rho+3\epsilon/4)}.$$

Recall that $\epsilon = (1-\rho)/2$, so that $\rho = 1-2\epsilon$ and that $\hat{\rho} = 1/(1+\epsilon/4)$. Using these substitutions, we obtain

$$1-\frac{s\hat{u}_n}{\mu} = 1-\frac{(1-2\epsilon)(1+\epsilon/4)}{1-2\epsilon+3\epsilon/4}$$

$$
\begin{aligned}
&= \frac{3\epsilon/4 - \epsilon/4 + 2\epsilon^2/4}{1 - 5\epsilon/4} \\
&= \frac{\epsilon(1+\epsilon)/2}{1-5\epsilon/4} \\
&\geq \frac{\epsilon}{2}.
\end{aligned}
$$

Q.E.D.

To obtain an upper bound for the probability in Eq. (82), we substitute Eqs. (83) and (85) into Eq. (82). Given the assumption that $s \leq \gamma n/\beta_n$, we have that

$$
\begin{aligned}
\mathbb{P}(|\mathcal{N}^*(S)| \leq \hat{u}_n s) &\leq \mathbb{P}\left(\text{Bino}\left(\beta_n s\,, \frac{b_n}{n}(\rho + 3\epsilon/4)\right) \leq \hat{u}_n s\right) \\
&\overset{(a)}{\leq} \exp\left(-\frac{1}{2}\left(\frac{\epsilon}{2}\right)^2 \mu\right) \\
&\overset{(b)}{=} \exp\left(-\frac{\epsilon^2}{8} \cdot \frac{80 \ln n}{\epsilon^2}(\rho + 3\epsilon/4)s\right) \\
&= \exp(-(10 \ln n)(\rho + 3\epsilon/4)s) \\
&\overset{(c)}{\leq} \exp(-(5 \ln n)s) \\
&= \frac{1}{n^{5s}}.
\end{aligned} \tag{86}
$$

for all sufficiently large n. Step (a) is based on a multiplicative form of the Chernoff bound (cf. Chapter 4 of [5]), $\mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{1}{2}\delta^2\mu\right)$, where X is a binomial random variable with $\mathbb{E}(X) = \mu$, and

$$
\delta = 1 - \frac{s\hat{u}_n}{\mu} \geq \epsilon/2, \tag{87}
$$

where the last inequality follows from Lemma A.4. Step (b) follows from Eq. (83), and (c) from the assumption that $\rho \geq 1/2$.

We now apply Eq. (86) to subsets of $\hat{\Gamma}$, and use the union bound. We have, for all sufficiently large n, that

$$
\begin{aligned}
\mathbb{P}\left(\hat{G} \text{ is not a } (\gamma/\hat{u}_n\,, \hat{u}_n)\text{-expander}\right) &\leq \mathbb{P}(\exists S \subset \hat{\Gamma} \text{ such that: } |S| \leq |\hat{\Gamma}|\gamma/\hat{u}_n \text{ and } |\mathcal{N}^*(S)| \leq \hat{u}_n|S|) \\
&\overset{(a)}{\leq} \sum_{s=1}^{|\hat{\Gamma}|\gamma/\hat{u}_n} \left(\sum_{S\subset\hat{\Gamma}, |S|=s} \mathbb{P}(|\mathcal{N}^*(S)| \leq \hat{u}_n s)\right) \\
&\leq \sum_{s=1}^{|\hat{\Gamma}|\gamma/\hat{u}_n} \binom{|\hat{\Gamma}|}{s} \mathbb{P}(|\mathcal{N}^*(S)| \leq \hat{u}_n s) \\
&\overset{(b)}{<} \sum_{s=1}^{|\hat{\Gamma}|\gamma/\hat{u}_n} b_n^s \mathbb{P}(|\mathcal{N}^*(S)| \leq \hat{u}_n s) \\
&\overset{(c)}{\leq} \sum_{s=1}^{|\hat{\Gamma}|\gamma/\hat{u}_n} b_n^s \frac{1}{n^{5s}}
\end{aligned}
$$

$$\leq \sum_{s=1}^{\infty} (b_n/n^5)^s$$
$$= \frac{b_n/n^5}{1 - b_n/n^5}. \tag{88}$$

Step (a) is the union bound. In step (b), we used the bound $\binom{n}{k} \leq n^k$, and the fact that $|\hat{\Gamma}| = m_n = \frac{\rho}{\hat{\rho}} b_n < b_n$. Step (c) follows from Eq. (86). Because $\beta_n \gg \ln n$, we have that $b_n \lesssim \frac{n \ln n}{\beta_n} \ll n$, and hence

$$\frac{b_n}{n^5} \leq \frac{1}{n^3}, \tag{89}$$

for all sufficiently large n. Combining Eqs. (88) and (89), we conclude that

$$\mathbb{P}\left(\hat{G} \text{ is not a } \left(\tfrac{\gamma}{\hat{u}_n}, \hat{u}_n\right)\text{-expander}\right) \leq \frac{1}{n^3}, \tag{90}$$

for all sufficiently large n. This proves our claim. Q.E.D.

Appendix B: Expanded Modular Architectures

In this appendix, we start by describing the graph product, and subsequently we discuss the implications of using an expander graph.

Construction of the Architecture. We first express the average degree as a product, $d_n = d_n^m \cdot d_n^e$, where the relative magnitudes of d_n^m and d_n^e are a design choice. The architecture is constructed as follows.

1. Similar to the case of the Modular architecture, partition I and J into equal-sized clusters of size d_n^m. We will refer to the index set of the queue and server clusters as $\mathcal{Q}$ and $\mathcal{S}$, respectively. For any $i \in I$ and $j \in J$, denote by $q(i) \in \mathcal{Q}$ and $s(j) \in \mathcal{S}$, the indices of the queue and server clusters to which i and j belong, respectively.
2. Let g_n^e be a bipartite graph of maximum degree d_n^e whose left and right nodes are the queue and server clusters, $\mathcal{Q}$ and $\mathcal{S}$, respectively. Let E^e be the set of edges of g_n^e.
3. To construct the interconnection topology $g_n = (I \cup J, E)$, let $(i,j) \in E$ if and only if their corresponding queue and server clusters are connected in g_n^e, i.e., if $(q(i), s(j)) \in E^e$.

Note that by the above construction, each queue is connected to at most d_n^e server clusters through g_n^e, and within each connected cluster, to d_n^m servers. Therefore, the maximum degree of g_n is $d_n^m \cdot d_n^e = d_n$.

Scheduling Policy. The scheduling policy requires the knowledge of the arrival rate vector, $\boldsymbol{\lambda}_n$, and involves two stages. For a given $\boldsymbol{\lambda}_n$, the computation in the first stage is performed only once, while the steps in the second stage are repeated throughout the operation of the system.

1. Compute a feasible flow, $\{f_{q,s}\}_{(q,s)\in E^e}$, over the graph g_n^e, where the incoming flow at each queue cluster $q \in \mathcal{Q}$ is equal to $\sum_{i\in q}\lambda_i$, and the outgoing flow at each server cluster $s \in \mathcal{S}$ is constrained to be less than or equal to $\frac{1+\rho}{2}d_n^m$. (It turns out that, under our assumptions, such a feasible flow exists [6].) Denote by $f_{q,s}$ the total rate of flow from the queue cluster q to the server cluster s.
2. Arriving jobs first wait in queue until they are fetched by a server. When a server becomes available, it chooses a neighboring queue cluster (w.r.t. the topology of g_n^e) with probability roughly proportional to the flow between the clusters. In particular, a server in cluster s chooses the queue cluster q with probability

$$p_{s,q} = \frac{f_{q,s}}{\sum_{q'\in\mathcal{N}(s)} f_{q',s}} \cdot \frac{1+\rho}{2} + \frac{1}{\deg(s)} \cdot \frac{1-\rho}{2}, \tag{91}$$

where $\deg(s)$ is the degree of s in g_n^e. Within the chosen cluster, the server starts serving a job from an arbitrary non-empty queue, or, if all queues in the cluster are empty, the server initiates an idling period whose length is exponentially distributed with mean 1.

When the graph g_n^e is an expander graph, we refer to the topology created via the above procedure as an *Expanded Modular architecture generated by* g_n^e.

Note that an Expanded Modular architecture is constructed as a "product" between an expander graph across the queue and server clusters, and a fully connected graph for each pair of connected clusters. As a result, its performance is also of a hybrid nature: the expansion properties of g_n^e guarantee a large capacity region, while a diminishing delay is obtained as a result of the growing size of the server and queue clusters. We summarize this in the next theorem. Here we assume that d_n^e is sufficiently large so that the expander graph described in Lemma 3.3 exists. The reader is referred to Section 3.4.5 of [6] for the proof of the theorem (although with different choices for some of the constants).

Theorem B.1 (Capacity and Delay of Expanded Modular Architectures) *Suppose that* $d_n = d_n^m \cdot d_n^e$. *Let* $\gamma = \sqrt{\rho}$ *and* $\beta_n = \frac{1}{2}\cdot\frac{\ln(1/\rho)}{1+\ln(1/\rho)}\ d_n^e$. *Let* g_n^e *be a* $(\gamma/\beta_n, \beta_n)$*-expander with maximum degree* d_n^e, *and let* g_n *be an Expanded Modular architecture generated by* g_n^e. *If*

$$u_n \le \frac{1+\rho}{2}\beta_n = \frac{1+\rho}{4}\cdot\frac{\ln(1/\rho)}{1+\ln(1/\rho)}\ d_n^e, \tag{92}$$

then, under the scheduling policy described above, we have that

$$\sup_{\boldsymbol{\lambda}_n\in\boldsymbol{\Lambda}_n(u_n)} \mathbb{E}\left(W|\boldsymbol{\lambda}_n\right) \lesssim \frac{c}{d_n^m}, \tag{93}$$

where c *is a constant that does not depend on* n.

A Tradeoff between the Size of the Capacity Region and the Delay. For the Expanded Modular architecture, the relative values of d_n^m and d_n^e reflect a design choice: a larger value of d_n^e ensures a larger capacity region, while a larger value of d_n^m yields smaller delays. Therefore, while the Expanded Modular architecture is able to provide a strong delay guarantee that applies to *all* arrival rate vectors in $\mathbf{\Lambda}_n(u_n)$, it comes at the expense of either a slower rate of diminishing delay (small d_n^m) or a smaller capacity region (small d_n^e).

References

[1] A. S. Asratian, T. M. J. Denley, and R. Haggkvist. *Bipartite Graphs and their Applications.* Cambridge University Press, 1998.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, 2012.

[3] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris. *Fundamentals of Queueing Theory.* John Wiley & Sons, 2008.

[4] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[5] M. Mitzenmacher and E. Upfal. *Probability and computing: Randomized Algorithms and Probabilistic Analysis.* Cambridge University Press, 2005.

[6] K. Xu. *On the power of (even a little) flexibility in dynamic resource allocation.* PhD thesis, Massachusetts Institute of Technology, 2014. Available at `http://hdl.handle.net/1721.1/91101`.