Noname manuscript No.
(will be inserted by the editor)

Asymptotic Independence of Servers' Activity in Queueing Systems with Limited Resource Pooling

Virag Shah · Gustavo de Veciana

Received: date / Accepted: date

Abstract We consider multi-class multi-server queuing systems where a subset of servers, called a server pool, may collaborate in serving jobs of a given class. The pools of servers associated with different classes may overlap, so the sharing of server resources across classes is done via a dynamic allocation policy based on a fairness criterion. We consider an asymptotic regime where the total load increases proportionally with the system size. We show that under limited scaling in size of server pools the stationary distribution for activity of a *fixed finite* subset of servers has asymptotically a product form, which in turn implies a concentration result for server activity. In particular, we establish a clear connection between the scaling of server pools' size and asymptotic independence. Further, these results are robust to the service requirement distribution of jobs.

For large-scale cloud systems where heterogeneous pools of servers collaborate in serving jobs of diverse classes, a concentration in server activity indicates that the overall power and network capacity that need to be provisioned may be substantially lower than the worst case, thus reducing costs.

Keywords Resource pooling · server activity · concentration · mean field · insensitivity · power · network capacity

Mathematics Subject Classification (2000) 60K25

V. Shah
The University of Texas at Austin
Department of ECE
Austin, Tx 78712 USA
E-mail: virag@utexas.edu

G. de Veciana
The University of Texas at Austin
Department of ECE
Austin, Tx 78712 USA
E-mail: gustavo@ece.utexas.edu

1 Introduction

Modern Internet services are increasingly driven by cloud infrastructure where thousands of collocated servers are employed in a centralized fashion to serve user requests. An important problem in engineering such large scale systems is the optimization and efficient use of resources. The design of such systems is made complex by the dynamic characteristics of service demands, which include stochastic arrivals of user requests/jobs, diversity in demand types, and random service requirements.

System designers often adopt a pessimistic approach towards resource allocation, in that, they aim for acceptable user-performance under extreme or even worst-case scenarios. However, such extreme scenarios may be unlikely (or may be made unlikely) and a pessimistic design may result in overprovisioning. A basic question in this setting is: For what system configurations and demand characteristics can we be optimistic in provisioning resources?

This paper has three key messages which we discuss below. The first message: *concentration in servers' activity facilitates resource provisioning.* As systems become large and service types become more diverse such that no single service dominates resource usage, the load across individual servers becomes increasingly uncorrelated. This may in turn result in concentration of servers' activity, i.e., the distribution of the number of active servers is concentrated around its mean. Such a result enables one to provision for the peak power capacity to be close to the average power requirement without a significant risk of overload. Similarly, for content delivery applications where activity of a server is connected to the rate at which bits are downloaded from the server, such concentration results would allow one to provision for a shared network link with capacity close to the average traffic demand without significantly affecting user-performance.

Existence of such a concentration result depends on the extent to which there is diversity and independence in the load spread across the servers. To better understand how diversity in service types impacts servers' activity, consider a system with m servers, each with service capacity μ. Let the job arrival rate be λm and mean service requirement of jobs be ν. For stability, assume $\lambda\nu < \mu$. We will consider systems that use resource pooling, in that the capacity of multiple servers may be pooled together as follows: if k servers are pooled together to serve a job, the job can be served at a maximum rate of $k\mu$. Note, however, these resources are shared among jobs and the pools may overlap. In this setting, consider the following two extreme cases:

Case 1: *Single service type and complete resource pooling:* Suppose that jobs belong to a single service type, and that all m servers can be pooled to serve each job. This system can be modeled as a $G/G/1$ queue with arrival rate λm and load $\rho = \lambda\nu/\mu$. For a work-conserving service policy, either all m servers are active or idle at the same time with probability ρ and $1 - \rho$ respectively.

Case 2: *Multiple service types and no resource pooling:* Now, suppose that there are m job classes. Each job class has a dedicated server. The arrivals and service requirements of different classes are independent. Suppose the arrival rate for each class is λ, and mean service requirement for jobs in each class is ν. This system can be modeled as consisting of m independent $G/G/1$ queues, each with load $\rho = \lambda\nu/\mu$. For queues with work conserving service policy, at any time t each server is active with probability ρ and the activities of different servers are independent. By Weak Law of Large Numbers, for any $\epsilon > 0$ the stationary probability that the number of active servers exceeds $(1+\epsilon)\rho m$ tends to 0 as $m \to \infty$.

In Case 2 the servers' activity concentrate due to independence in load, thus facilitating resource provisioning for the large scale system. By contrast, in Case 1 the activities of different servers are correlated due to complete resource pooling and one may need to provision for the peak number of servers being active. Thus, a question arises: do servers' activity concentrate in systems where limited resource pooling is allowed? Such systems fall in between the above two extreme cases, in that, there may be diverse service types and a limited amount of resource pooling which correlates instantaneous server activities.

The second message: *servers' activity concentrate even if we allow limited resource pooling of servers.* To better understand the impact of limited resource pooling, in this paper we consider multi-class multi-server systems where for each job class the capacity of a unique subset of servers can be pooled to jointly serve the class's jobs. Furthermore the pools of servers serving different classes may overlap, which opens up an opportunity to dynamically vary the allocation of service capacity across job classes. We consider the case where the service rate allocated to each class depends on the numbers of jobs across classes, and call the corresponding policy a resource allocation policy.

Such a system can model a centralized content delivery infrastructure where each file is replicated across multiple servers so as to address high demands and possible reliability issues. Systems which combine multipath transport with server diversity can support parallel downloads from multiple servers, where different chunks of a file can be downloaded in parallel from servers possibly via multiple paths. The resource allocation policy would thus model the dynamically varying sum-rate a download job receives from its server pool.

In this paper, we focus on Balanced Fair (BF) resource allocation, which was introduced in [4] as a bandwidth sharing model for networks where flows compete for the available bandwidth across network links on their pre-defined route. Several studies have shown close structural relationship between balanced fairness and proportional fairness [3, 4, 13]. Further, equivalence and comparison results among several fairness policies in [16, 17] imply that, for large systems employing resource pooling, the choice among fairness principles in resource allocation and performance may be of limited significance.

To understand the role of overlapping pools on possible concentration properties of such systems, we consider a sequence of systems where the number of servers m grows. We allow total system load and total server capacity to scale linearly with m. For a given m we consider server pools of fixed size $c^{(m)}$, which may scale with m but as $o(m)$. We assume that the load across different server pools (equivalently, job classes) is homogeneous. This may be achieved by, for example, grouping of several service types into a class so that the overall load per group is roughly the same (see [16,17] for more discussions on impact of heterogeneity in loads). For such a system, we show that the joint stationary distribution of the activity of a *fixed finite* subset of servers takes a product form as $m \to \infty$, which in turn implies that a WLLN holds for the servers' activity. In summary, as long as resource pools are of size $o(m)$ one will see a concentration in server activity.

The above concentration result is 'insensitive', in that, the dependence on the service requirement distribution for each class is only through its mean. This follows from our adoption of insensitive balanced fair resource allocation [4]. This brings us to the third message: *one's optimism in resource allocation due to concentration in servers' activity is independent of service requirement distribution, i.e., only depends on its mean.* This is analogous to insensitivity in symmetric queueing systems where the distribution of the number of active jobs in a system is known to be insensitive to service-time distributions [10], although our interest is mainly in the distribution of servers' activity for large scale coupled systems.

Relationship to prior work: Prior work which is closest in spirit to ours is that studying the existence of a mean field regime for the super-market queuing model [7, 14, 18]. In the super-market model the servers are coupled through a routing policy, unlike our model where they are coupled through a servicing policy. In the supermarket model, upon arrival of a job a random subset of servers of size d is selected and the job is routed to the server with the least number of jobs waiting for service. For a fixed value of d, asymptotic independence in the number of waiting jobs for a fixed finite subset of servers was shown in [7] for several classes of service distributions.

A mean field result for the number of waiting jobs was also shown for a symmetric loss network model [11, 19], where upon arrival of a job (or a call in their terminology) it is allocated to a fixed w number of servers at random. In this work, rather than routing to one of the w servers as in supermarket model, each job 'locks' resources at w servers for a random time. The maximum number of jobs that can lock resources at a given server at any point in time is fixed. Again, w is assumed to be constant. Further, the random locking time is assumed exponential.

In comparison, in this paper we consider a setting where a job arrives with a random service requirement, is served jointly by a subset of servers, and leaves the system upon completion of its service. Sojourn times of jobs thus depend on how server resources are shared across different job-types. We allow the number of servers that can be pooled together for serving a job to scale with m. We also let the distribution of the service requirement be arbitrary.

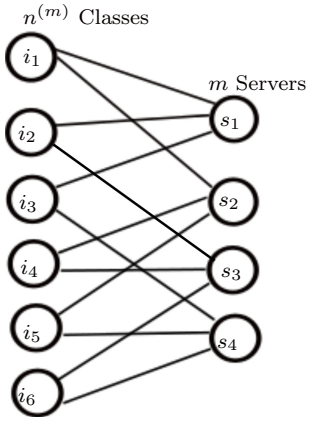

Fig. 1: Graph $\mathcal{G}^{(m)} = (F^{(m)} \cup S^{(m)}; E^{(m)})$ for $m = 4$ and $c^{(m)} = 2$ modeling availability of a servers $S^{(m)}$ to serve jobs in classes $F^{(m)}$.

Under these assumptions, we are able to show the existence of a mean field only for servers' activity, and not for the number of waiting jobs.

In terms of tools used, instead of analyzing sample paths of the underlying stochastic processes as done in above prior work, we use the knowledge of stationary distribution of waiting jobs under balanced fair resource allocation [4]. Since its proposal in [4] as a bandwidth sharing policy for wireline network, it has been a useful device towards analyzing user-performance in several kinds of network models [3, 5, 6, 16, 17].

Outline of the paper: In Section 2 we describe our system model. In Section 3 we provide our main result where we show concentration in servers' activity. In Section 4 we consider engineering implications of this result.

2 System Model

Consider a system with a set $S^{(m)} = \{s_1, s_2, \ldots, s_m\}$ of m servers. Each server has service capacity $\xi > 0$. Jobs arrive into the system as an independent Poisson process with rate λm. Job service requirements are i.i.d. with mean ν. Let $\rho = \lambda\nu$. We assume that $\rho < \xi$ to ensure stability. Upon arrival of a job, $c^{(m)} > 1$ servers are chosen at random, and their capacity pooled, to serve this job. Let $F^{(m)}$ represent the set of all possible pools of size $c^{(m)}$. Let $n^{(m)} \triangleq |F^{(m)}|$. Thus, $n^{(m)} = \binom{m}{c^{(m)}}$.

We view the system as consisting of $n^{(m)}$ job classes, where arrivals for each class occur as an independent Poisson process with rate $\lambda m/n^{(m)}$. Let $\boldsymbol{\rho}^{(m)} = (\rho_i^{(m)} : i \in F^{(m)})$, where $\rho_i^{(m)} = \rho m/n^{(m)}$ denotes the load associated with class i.[1] We view the association of classes with server pools via a bipartite

[1] Our model may be generalized in the following ways without affecting our results:
1) The service requirement distribution may be different for each class as long as the mean

graph $\mathcal{G}^{(m)} = (F^{(m)} \cup S^{(m)}; E^{(m)})$ where an edge $e \in E^{(m)}$ exists if it connects a class $i \in F^{(m)}$ to a server $s \in S^{(m)}$ associated with the server pool of i, see Fig. 1. For each class $i \in F^{(m)}$, let $N_i^{(m)}$ denote its neighbors, i.e., the set of servers available to serve jobs in class i.

Let $q_i(t)$ denote the *set* of ongoing jobs of class i at time t, i.e., jobs which have arrived but have not completed service. Let $\mathbf{x}(t) = (x_i(t) : i \in F^{(m)})$, where $x_i(t) \triangleq |q_i(t)|$, i.e., $\mathbf{x}(t)$ captures the *number of ongoing jobs in each class*. Let $\mathbf{X}^{(m)}(t)$ be the corresponding random vector capturing the random number of ongoing jobs in each class at time t.

For any $\mathbf{x}(t)$, let $A_{\mathbf{x}(t)}$ denote the set of active classes, i.e., classes with at least one ongoing job. Further, for each $s \in S^{(m)}$, let

$$Y_s^{(m)}(t) = \mathbf{1}_{\left\{\exists i \in A_{\mathbf{x}(t)} \text{ s.t. } s \in N_i^{(m)}\right\}}.$$

If $Y_s^{(m)}(t)$ is 1 we say that the server s is active at time t.

For each $v \in q_i(t)$ and $s \in S^{(m)}$, let $b_{v,s}(t)$ be the rate at which server s serves job v at time t. Let $b_v(t)$ be the total rate at which job v is served at time t by the file-server system. If job v arrives at time t_v^a and has service requirement η_v, then it departs at time t_v^d such that $\eta_v = \int_{t_v^a}^{t_v^d} b_v(t)dt$.

Our service model is subject to the following assumption.

Assumption 1 *Sharing of system service capacity among ongoing jobs is such that:*

1. *A server s can concurrently serve multiple jobs as long as $\sum_v b_{v,s}(t) \leq \xi$ for all t.*
2. *Multiple servers can concurrently serve a job v at time t giving a total service rate $b_v(t) = \sum_s b_{v,s}(t)$.*
3. *The service rate $b_{v,s}(t)$ allocated to a job v at server s at time t depends only on its job's class and the numbers of ongoing jobs $\mathbf{x}(t)$. Thus, for each i, the jobs in $q_i(t)$ receive equal rate at time t which depends only on $\mathbf{x}(t)$.*

Let $r_i^{(m)}(\mathbf{x}')$ be the total rate at which class i jobs are served at time t when $\mathbf{x}(t) = \mathbf{x}'$, i.e., at any time t, $r_i^{(m)}(\mathbf{x}(t)) = \sum_{v \in q_i(t)} b_v(t)$. Let $\mathbf{r}^{(m)}(\mathbf{x}) = (r_i^{(m)}(\mathbf{x}) : i \in F^{(m)})$. We call the vector function $\mathbf{r}^{(m)}(.)$ *the resource allocation.*

Polymatroid Capacity Region: For a given graph $\mathcal{G}^{(m)}$, it was shown that under Assumption 1 the set of feasible rate vectors, i.e., the *capacity region*, is a polymatroid, see [16]. A polytope $\mathcal{C}$ is a *polymatroid* if there exists a set function $\mu : \{0,1\}^F \to \mathbb{R}$ such that

$$\mathcal{C} = \left\{ \mathbf{r} \geq \mathbf{0} : \sum_{i \in A} r_i \leq \mu(A),\ \forall A \subset F \right\},$$

service requirement is same for each class.

2) The arrival rate and mean service requirement may be different for each class as long as their product (which equals to $\rho_i^{(m)}$) is same for each class.

and if $\mu(.)$ satisfies the following properties:
1) Normalized: $\mu(\emptyset) = 0$.
2) Monotonic: if $A \subset B$, $\mu(A) \leq \mu(B)$.
3) Submodular: for all $A, B \subset F$,

$$\mu(A) + \mu(B) \geq \mu(A \cup B) + \mu(A \cap B).$$

A function $\mu(.)$ satisfying the above properties is called a *rank function*. Polymatroids and submodular functions are well studied in the literature, see e.g., [8, 15].

The following proposition holds under Assumption 1.

Proposition 1 ([16]) *Consider graph $\mathcal{G}^{(m)}$. For each $A \subset F^{(m)}$, let*

$$\mu^{(m)}(A) \triangleq \xi \left| \cup_{i \in A} N_i^{(m)} \right|,$$

i.e., $\mu^{(m)}(A)$ is the maximum sum rate at which the jobs in classes belonging to set A can be served. Further, let

$$\mathcal{C}^{(m)} \triangleq \left\{ \mathbf{r} \geq \mathbf{0} : \sum_{i \in A} r_i \leq \mu^{(m)}(A),\ \forall A \subset F^{(m)} \right\}.$$

Then, the following statements hold:
1) $\mu^{(m)}(.)$ is a submodular function.
2) Under Assumption 1, $\mathcal{C}^{(m)}$ is the polymatroid capacity region associated with the system.

Further, the following holds for the rank function $\mu^{(m)}(.)$.

Proposition 2 *For each $k \leq n^{(m)}$, we have*

$$\sum_{A \subset F^{(m)}: |A| = k} \mu^{(m)}(A) = \xi m \left(\binom{n^{(m)}}{k} - \binom{n^{(m-1)}}{k} \right).$$

The proof of this proposition is straightforward. Notice that the term $\binom{n^{(m)}}{k} - \binom{n^{(m-1)}}{k}$ captures the number of subsets of $F^{(m)}$ of size k which are served by a given server.

Given the capacity region $\mathcal{C}^{(m)}$, there are several feasible resource allocation policies $\mathbf{r}^{(m)}(.)$ which may dynamically vary the service rate of each class $\mathbf{r}^{(m)}(\mathbf{x}) \in \mathcal{C}^{(m)}$ for each $\mathbf{x}$, to achieve fairness/load-balancing objectives. For a discussion and comparison of different resource allocation policies, see [17]. In this paper, we assume Balanced Fair resource allocation leveraging its analytical tractability and insensitivity properties. Further, it serves as an approximation to proportionally fair/α-fair allocation policies on polymatroid capacity regions [3, 4, 13]

Balanced fair resource allocation: Balanced Fairness (BF) was introduced in [4] as a bandwidth allocation policy in networks to provide insensitivity in stationary distribution for the number of jobs in a bandwidth sharing

network. We adopt balanced fairness to provide a resource allocation policy for our system with capacity region $\mathcal{C}^{(m)}$.

Formally, the balanced fair resource allocation for a polymatroid capacity region can be given as follows, see [4]. Let $\mathbf{e}_i$ be a unit vector in i^{th} direction. For each $i \in F^{(m)}$ and for each $\mathbf{x}$ we have

$$r_i^{(m)}(\mathbf{x}) = \frac{\Phi^{(m)}(\mathbf{x} - \mathbf{e}_i)}{\Phi^{(m)}(\mathbf{x})}, \tag{1}$$

where the function $\Phi^{(m)}$ is called a balance function and is defined recursively as follows: $\Phi^{(m)}(\mathbf{0}) = 1$, and $\Phi^{(m)}(\mathbf{x}) = 0$ $\forall \mathbf{x}$ s.t. $x_i < 0$ for some i, otherwise,

$$\Phi^{(m)}(\mathbf{x}) = \max_{A \subset A_{\mathbf{x}}} \left\{ \frac{\sum_{i \in A} \Phi^{(m)}(\mathbf{x} - \mathbf{e}_i)}{\mu^{(m)}(A)} \right\}. \tag{2}$$

As shown in [4], (1) ensures insensitivity, while (2) ensures that $\mathbf{r}^{(m)}(\mathbf{x})$ for each $\mathbf{x}$ lies in the capacity region, i.e., the constraints $\sum_{i \in A} r_i^{(m)}(\mathbf{x}) \leq \mu^{(m)}(A)$ are satisfied for each A. It also ensures that there exists a set $B \subset A_{\mathbf{x}}$ for which we have $\sum_{i \in B} r_i^{(m)}(\mathbf{x}) = \mu^{(m)}(B)$. In fact the balanced fair resource allocation is the unique policy satisfying the above properties.

Since $\rho < \xi$, one can check that $\boldsymbol{\rho}^{(m)}$ lies in the interior of $\mathcal{C}^{(m)}$. It follows from a stability result in [4] that the process $(\mathbf{X}^{(m)}(t) : t \in \mathbb{R})$ is stationary and ergodic under balanced fair resource allocation. Further, the stationary distribution is given by

$$\pi^{(m)}(\mathbf{x}) = \frac{\Phi^{(m)}(\mathbf{x})}{G^{(m)}(\boldsymbol{\rho}^{(m)})} \prod_{i \in A_{\mathbf{x}}} \left(\rho_i^{(m)}\right)^{x_i}, \tag{3}$$

where,

$$G^{(m)}(\boldsymbol{\rho}^{(m)}) = \sum_{\mathbf{x}'} \Phi^{(m)}(\mathbf{x}') \prod_{i \in A_{x'}} \left(\rho_i^{(m)}\right)^{x_i'}.$$

Recall, $\rho_i^{(m)} = \rho m / n^{(m)}$ for each i.

The existence of such an expression for stationary distribution makes balanced fairness amenable to our analysis. While, in general, BF may result in wasteful resource allocation, e.g., BF is not Pareto efficient for certain triangle networks studied in [4], for polymatroid capacity regions BF is has been shown to be Pareto efficient:

Proposition 3 ([16]) *For polymatroid capacity regions, the balanced fair resource allocation is Pareto efficient, i.e.,* $\sum_{i \in A_{\mathbf{x}}} r_i^{(m)}(\mathbf{x}) = \mu^{(m)}(A_{\mathbf{x}})$ *for each* $\mathbf{x}$.

Thus, the total service rate is same for all $\mathbf{x}$ such that $A_{\mathbf{x}} = A$. Let

$$G_A^{(m)}(\boldsymbol{\rho}^{(m)}) \triangleq \sum_{\mathbf{x}: A_{\mathbf{x}} = A} \Phi^{(m)}(\mathbf{x}) \prod_{i \in A} \left(\rho_i^{(m)}\right)^{x_i}.$$

Then, under the stationary distribution $\pi^{(m)}$, for each $A \subset F^{(m)}$ we have

$$Pr_{\pi^{(m)}}\left(A_{\mathbf{X}^{(m)}} = A\right) = \frac{G_A^{(m)}(\boldsymbol{\rho}^{(m)})}{G^{(m)}(\boldsymbol{\rho}^{(m)})}.$$

Using Proposition 3 one can obtain a recursive expression for $G_A^{(m)}(\boldsymbol{\rho})$ as given below. For a proof, see [16].

Proposition 4 ([16]) *Let $G_\emptyset^{(m)}(\boldsymbol{\rho}^{(m)}) = 1$. Then, $G_A^{(m)}(\boldsymbol{\rho}^{(m)})$ for each $A \subset F^{(m)}$ can be computed recursively as*

$$G_A^{(m)}(\boldsymbol{\rho}^{(m)}) = \frac{\sum_{i\in A}\rho_i^{(m)} G_{A\setminus\{i\}}^{(m)}(\boldsymbol{\rho})}{\mu^{(m)}(A) - \sum_{j\in A}\rho_j^{(m)}}. \tag{4}$$

3 Asymptotic independence and concentration in servers' activity

The following lemma provides the expectation of the overall servers' activity

Lemma 1 *For a given m, if $\rho < \xi$ then we have*

$$E_{\pi^{(m)}}\left[\mu^{(m)}(A_{\mathbf{X}^{(m)}})\right] = \rho m, \tag{5}$$

Proof By definition of $\mu^{(m)}(.)$ and Assumption 1 we have

$$\frac{1}{t}\int_0^t \mu(A_{x(\tau)})d\tau = \frac{1}{t}\int_0^t \sum_s \sum_v b_{v,s}(\tau)d\tau = \frac{1}{t}\int_0^t \sum_v b_v(\tau)d\tau$$

Further, by ergodicity of the system we have

$$E_{\pi^{(m)}}\left[\mu^{(m)}(A_{\mathbf{X}^{(m)}})\right] = \frac{1}{t}\int_0^t \mu(A_{X(\tau)})d\tau \text{ a.s.}$$

Since service requirements of jobs are i.i.d. and the system is ergodic, for almost every sample path the term $\frac{1}{t}\int_0^t \sum_v b_v(\tau)d\tau$ tends to $\lambda\nu m$. Hence the result holds. □

Recall, $Y_s^{(m)}$ captures the activity of server s. By Pareto optimality of the balanced fair resource allocation for our system we have

$$E_{\pi^{(m)}}\left[\mu^{(m)}(A_{\mathbf{X}^{(m)}})\right] = \xi E_{\pi^{(m)}}\left[\sum_{s\in S^{(m)}} Y_s^{(m)}\right].$$

By symmetry and linearity of expectation, for each $s \in S^{(m)}$ we have

$$E_{\pi^{(m)}}[Y_s^{(m)}] = \rho/\xi.$$

Indeed, showing concentration in $\sum_{l=1}^{m} Y_{s_l}^{(m)}$ as $m \to \infty$ is equivalent to showing concentration in $\mu^{(m)}(A_{\mathbf{X}^{(m)}})$ close to its mean. Further, for a given m, $\left(Y_{s_1}^{(m)}, Y_{s_2}^{(m)}, \ldots, Y_{s_m}^{(m)}\right)$ is an exchangeable vector of random variables. A weak convergence result for a sequence of exchangeable vectors was shown in [1,9], which when applied to $\left(\left(Y_{s_1}^{(m)}, Y_{s_2}^{(m)}, \ldots, Y_{s_m}^{(m)}\right) : m \in \mathbb{N}\right)$ implies that $\frac{1}{m}\sum_{l=1}^{m} Y_{s_l}^{(m)}$ converges to a constant in probability if and only if the joint-distribution of $\left(Y_{s_1}^{(m)}, Y_{s_2}^{(m)}, \ldots, Y_{s_k}^{(m)}\right)$ for a finite k takes a product form as $m \to \infty$. The following theorem is the first main result in this paper.

Theorem 1 *Consider a sequence of systems with an increasing number of servers m. Suppose that the total arrival rate of jobs is λm for the m^{th} system, and that the service capacity of each server is a constant ξ. Let load per server $\rho = \lambda\nu$ be a constant such that $\rho < \xi$.*

Upon arrival of a job, $c^{(m)} > 1$ servers are selected at random for its service (equivalently, its class is selected at random). Assume $c^{(m)}$ is $o(m)$. Jobs share the server resources according to the balanced fair resource allocation. For each m, the system is stationary. Under stationary distribution, let $Y_s^{(m)}$ represent instantaneous activity of server s.

Then, the following equivalent statements hold:

(a) For any finite integer k, the random variables $Y_{s_1}^{(m)}, Y_{s_2}^{(m)}, \ldots, Y_{s_k}^{(m)}$ are asymptotically i.i.d. as $m \to \infty$.

(b) $\lim_{m\to\infty} \dfrac{\sum_{l=1}^{m} Y_{s_l}^{(m)}}{m} = \rho/\xi$ *in probability.*

Proof: Equivalence of *(a)* and *(b)* thus follows from Proposition 7.20 in [1]. We prove *(a)* below for $\xi = 1$ without loss of generality. Again by Proposition 7.20 in [1], it is sufficient to show that the result holds for $k = 2$.

Let

$$\mathcal{T}_{s,1}^{(m)} \triangleq \{A \subset F^{(m)} : s \in \cup_{i\in A} N_i^{(m)}\}$$

and similarly,

$$\mathcal{T}_{s,0}^{(m)} \triangleq \{A \subset F^{(m)} : s \notin \cup_{i\in A} N_i^{(m)}\}$$

Recall the definitions of $G_A^{(m)}(\boldsymbol{\rho}^{(m)})$ and $G^{(m)}(\boldsymbol{\rho}^{(m)})$. Then, for $b \in \{0, 1\}$, we have

$$\begin{aligned} Pr\left(Y_s^{(m)} = b\right) &= \sum_{\mathbf{x} \text{ s.t. } A_{\mathbf{x}} \in \mathcal{T}_{s,b}^{(m)}} \pi^{(m)}(\mathbf{x}) \\ &= \sum_{A \in \mathcal{T}_{s,b}^{(m)}} \frac{G_A^{(m)}(\boldsymbol{\rho}^{(m)})}{G^{(m)}(\boldsymbol{\rho}^{(m)})} \\ &= (1-\rho)\mathbf{1}_{\{b=0\}} + \rho\mathbf{1}_{\{b=1\}}. \end{aligned} \tag{6}$$

Further,

$$
\begin{aligned}
Pr\left(Y_{s_1}^{(m)} = b_1, Y_{s_2}^{(m)} = 0\right) &= \sum_{A \in \mathcal{T}_{s_1,b_1}^{(m)} \cap \mathcal{T}_{s_2,0}^{(m)}} \frac{G_A^{(m)}(\boldsymbol{\rho}^{(m)})}{G^{(m)}(\boldsymbol{\rho}^{(m)})} \\
&= \frac{\sum_{A \in \mathcal{T}_{s_1,b_1}^{(m)} \cap \mathcal{T}_{s_2,0}^{(m)}} G_A^{(m)}(\boldsymbol{\rho}^{(m)})}{\sum_{A \in \mathcal{T}_{s_2,0}^{(m)}} G_A^{(m)}(\boldsymbol{\rho}^{(m)})} \frac{\sum_{A \in \mathcal{T}_{s_2,0}^{(m)}} G_A^{(m)}(\boldsymbol{\rho}^{(m)})}{G^{(m)}(\boldsymbol{\rho}^{(m)})} \\
&= \frac{\sum_{A \in \mathcal{T}_{s_1,b_1}^{(m)} \cap \mathcal{T}_{s_2,0}^{(m)}} G_A^{(m)}(\boldsymbol{\rho}^{(m)})}{\sum_{A \in \mathcal{T}_{s_2,0}^{(m)}} G_A^{(m)}(\boldsymbol{\rho}^{(m)})} Pr\left(Y_{s_2}^{(m)} = 0\right) \\
&= \frac{\sum_{A \in \mathcal{T}_{s_1,b_1}^{(m)} \cap \mathcal{T}_{s_2,0}^{(m)}} G_A^{(m)}(\boldsymbol{\rho}^{(m)})}{\sum_{A \in \mathcal{T}_{s_2,0}^{(m)}} G_A^{(m)}(\boldsymbol{\rho}^{(m)})} (1-\rho) \qquad (7)
\end{aligned}
$$

Consider the denominator $\sum_{A \in \mathcal{T}_{s_2,0}^{(m)}} G_A^{(m)}(\boldsymbol{\rho}^{(m)})$. By symmetry, $\sum_{A \in \mathcal{T}_{s_2,0}^{(m)}} G_A^{(m)}(\boldsymbol{\rho}^{(m)}) = \sum_{A \in \mathcal{T}_{s_m,0}^{(m)}} G_A^{(m)}(\boldsymbol{\rho}^{(m)})$. Also for each $A \in \mathcal{T}_{s_m,0}^{(m)}$,

$$
\begin{aligned}
G_A^{(m)}(\boldsymbol{\rho}^{(m)}) &= \sum_{\mathbf{x}: A_{\mathbf{x}} = A} \Phi^{(m)}(\mathbf{x}) \left(\frac{m}{n^{(m)}} \rho\right)^{|\mathbf{x}|} \\
&= \sum_{\mathbf{x}: A_{\mathbf{x}} = A} \Phi^{(m)}(\mathbf{x}) \left(\frac{m-1}{n^{(m-1)}} \frac{(m-c^{(m)})\rho}{m-1}\right)^{|\mathbf{x}|},
\end{aligned}
$$

since $\frac{m-1}{n^{(m-1)}} \frac{m-c^{(m)}}{m-1} = \frac{m}{n^{(m)}}$.

However, it is easy to check that $\mathcal{T}_{s_m,0}^{(m)}$ is the power set of $F^{(m-1)}$. Thus,

$$
\begin{aligned}
\sum_{A \in \mathcal{T}_{s_m,0}^{(m)}} G_A^{(m)}(\boldsymbol{\rho}^{(m)}) &= \sum_{A \subset F^{(m-1)}} \sum_{\mathbf{x}: A_{\mathbf{x}} = A} \Phi^{(m-1)}(\mathbf{x}) \left(\frac{m-1}{n^{(m-1)}} \frac{(m-c^{(m)})\rho}{m-1}\right)^{|\mathbf{x}|} \\
&= \sum_{A \subset F^{(m-1)}} G_A^{(m-1)}\left(\boldsymbol{\rho}'^{(m-1)}\right),
\end{aligned}
$$

where,

$$
\boldsymbol{\rho}'^{(m-1)} \triangleq \left(\rho_i'^{(m-1)} \triangleq \frac{m-1}{n^{(m-1)}} \frac{(m-c^{(m)})\rho}{m-1} : i \in F^{(m-1)}\right).
$$

Thus, in turn,

$$
\begin{aligned}
\sum_{A \in \mathcal{T}_{s_2,0}^{(m)}} G_A^{(m)}(\boldsymbol{\rho}^{(m)}) &= \sum_{A \subset F^{(m-1)}} G_A^{(m-1)}\left(\boldsymbol{\rho}'^{(m-1)}\right) \\
&= G^{(m-1)}\left(\boldsymbol{\rho}'^{(m-1)}\right)
\end{aligned}
$$

Using similar arguments, one can show that

$$\sum_{A\in\mathcal{T}^{(m)}_{s_1,b_1}\cap\mathcal{T}^{(m)}_{s_2,0}} G^{(m)}_A(\boldsymbol{\rho}^{(m)}) = \sum_{A\in\mathcal{T}^{(m-1)}_{s_1,b_1}} G^{(m-1)}_A\left(\boldsymbol{\rho}'^{(m-1)}\right)$$

Combining above equalities, we get,

$$\begin{aligned}\frac{\sum_{A\in\mathcal{T}^{(m)}_{s_1,b_1}\cap\mathcal{T}^{(m)}_{s_2,0}} G^{(m)}_A(\boldsymbol{\rho}^{(m)})}{\sum_{A\in\mathcal{T}^{(m)}_{s_2,0}} G^{(m)}_A(\boldsymbol{\rho}^{(m)})} &= \frac{\sum_{A\in\mathcal{T}^{(m-1)}_{s_1,b_1}} G^{(m-1)}_A\left(\boldsymbol{\rho}'^{(m-1)}\right)}{G^{(m-1)}\left(\boldsymbol{\rho}'^{(m-1)}\right)}\\ &= \left(1-\frac{(m-c^{(m)})\rho}{m-1}\right)\mathbf{1}_{\{b_1=0\}} + \frac{(m-c^{(m)})\rho}{m-1}\mathbf{1}_{\{b_1=1\}}\end{aligned}$$

where the last equality follows from (6). By substituting this in (7) we get

$$\begin{aligned}&Pr\left(Y^{(m)}_{s_1}=b_1, Y^{(m)}_{s_2}=0\right)\\ &\quad= (1-\rho)\left(\left(1-\frac{(m-c^{(m)})\rho}{m-1}\right)\mathbf{1}_{\{b_1=0\}} + \frac{(m-c^{(m)})\rho}{m-1}\mathbf{1}_{\{b_1=1\}}\right) \quad (8)\end{aligned}$$

By using law of total probability, and taking the limit as $m\to\infty$, we get,

$$Pr\left(Y^{(m)}_{s_1}=b_1, Y^{(m)}_{s_2}=b_2\right) \overset{m\to\infty}{\longrightarrow} \prod_{i=1}^{2}\left((1-\rho)\mathbf{1}_{\{b_i=0\}} + \rho\mathbf{1}_{\{b_i=1\}}\right).$$

Hence the theorem holds. □

We now illustrate the impact of increasing the values of m and $c^{(m)}$ on the degree of pairwise independence of server activity by providing an explicit expression for Pearson's correlation coefficient for the activities of two servers as well as the total variation distance between their joint distribution and the associated product form distribution for finite m. Recall that for exchangeable random variables pairwise independence is sufficient for concentration. Let σ_Y represent variance of random variable Y and let $\mathrm{Cov}(Y,Z)$ represent covariance of random variables Y and Z. Using (8) we can show that the Pearson's correlation coefficient for the activities of two servers can be given as

$$\rho_{Y^{(m)}_{s_1}Y^{(m)}_{s_2}} \triangleq \frac{\mathrm{Cov}(Y^{(m)}_{s_1}, Y^{(m)}_{s_2})}{\sigma_{Y^{(m)}_{s_1}}\sigma_{Y^{(m)}_{s_1}}} = \frac{c^{(m)}}{m}.$$

The total variation distance between two probability measures P and P' on sigma algebra $\mathcal{F}$ on a sample space Ω is defined as

$$\delta(P,P') \triangleq \sup_{A\in\mathcal{F}} |P(A)-P'(A)|.$$

If Ω is finite then it can be equivalently given as

$$\delta(P, P') = \frac{1}{2} \sum_{\omega \in \Omega} |P(\omega) - P'(\omega)|.$$

Let $P_{Y_s^{(m)}}$ represent the distribution of $Y_s^{(m)}$ and let $P_{Y_{s_1}^{(m)} Y_{s_2}^{(m)}}$ represent the joint distribution of $Y_{s_1}^{(m)}$ and $Y_{s_2}^{(m)}$. Using (8) and direct computations we can show that

$$\delta\left(P_{Y_{s_1}^{(m)} Y_{s_2}^{(m)}}, P_{Y_{s_1}^{(m)}} P_{Y_{s_2}^{(m)}}\right) = 2\frac{c^{(m)}}{m}\rho(1-\rho).$$

We have assumed that the sizes of resource pools are homogeneous for a given system to obtain clean results. We conjecture that if pool sizes are heterogeneous but $o(m)$ then with sufficient diversity (randomness) in the choice of resource pools our concentration results will hold.

4 Engineering Implications

In previous section, we showed that the total number of active servers concentrates close to its mean. In this section we study its implication for provisioning of the peak power capacity and/or of a shared network link for such large scale systems.

Several modern systems are designed so that the power usage of a server is low when it is inactive [2,12]. Thus, the total instantaneous power draw in such systems is an increasing function of the total number of active servers. Thus, a concentration in servers' activity implies that the peak power draw is unlikely to be significantly away from the mean power consumption. This allows one to reduce infrastructure costs by provisioning for a peak power capacity which is close to the average power requirement without a significant risk of overload.

Similarly, for centralized content delivery systems where the servers are collocated and are connected to the users via a shared network link, see Fig. 2, a concentration in servers' activity facilitates provisioning of the network link. In such systems, the total number of active servers is proportional to the overall network traffic volume. Intuitively, Theorem 1 implies that if the bandwidth of the shared network link is greater than $(1+\epsilon)\rho m$, the link may cease to become a bottleneck as m becomes large. Below, we make this intuition more precise.

Consider a content delivery system where the bandwidth the shared network link is $\beta^{(m)}$. Thus, the maximum sum-rate at which bits may be downloaded from the servers is $\beta^{(m)}$. In this section we assume that if upon arrival of a job the service capacity $\mu^{(m)}\left(A_{\hat{\mathbf{X}}^{(m)}}(t)\right)$ exceeds $\beta^{(m)}$ then the job is blocked, where $\hat{\mathbf{X}}^{(m)}(t)$ represents the number of jobs in the modified system. Thus, $\hat{\mathbf{X}}^{(m)}(t)$ is restricted to remain within the following set:

$$\mathcal{A}^{(m)} \triangleq \left\{\mathbf{x} : \mu^{(m)}(A_{\mathbf{x}}) \leq \beta^{(m)}\right\}.$$

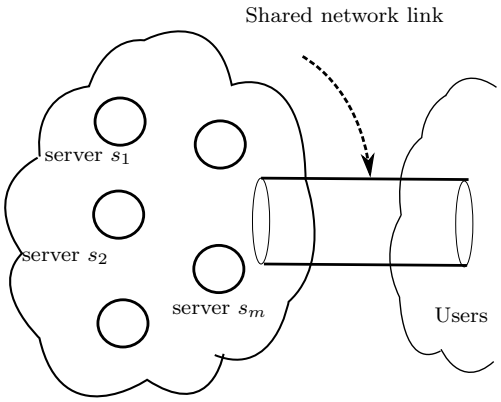

Fig. 2: A centralized content delivery system where collocated servers are connected to users via a shared network link.

For such a system, the stationary distribution $\hat{\mathbf{X}}^{(m)}$, namely $\hat{\pi}^{(m)}(.)$, is a truncated version of $\pi^{(m)}(.)$ (see (3) for definition of $\pi^{(m)}(.)$). Indeed this follows since $\pi^{(m)}(.)$ satisfies detailed balance conditions, see [10]. Thus, $\hat{\pi}^{(m)}(.)$ can be given as follows: for each $\mathbf{x}$,

$$\hat{\pi}^{(m)}(\mathbf{x}) = \frac{\pi^{(m)}(\mathbf{x})\mathbf{1}_{\{\mathbf{x}\in\mathcal{A}^{(m)}\}}}{\sum_{\mathbf{x}\in\mathcal{A}^{(m)}}\pi^{(m)}(\mathbf{x})}. \tag{9}$$

A class i arrival gets blocked if it sees a state $\mathbf{x}$ in the following set:

$$\mathcal{B}_i \triangleq \{\mathbf{x}\in\mathcal{A}^{(m)} : \mathbf{x}+\mathbf{e}_i \notin \mathcal{A}^{(m)}\}.$$

By PASTA, the probability that a class i arrival gets blocked is given by:

$$p_i^{(m)} = \sum_{\mathbf{x}\in\mathcal{B}_i}\hat{\pi}^{(m)}(\mathbf{x}).$$

Let

$$\mathcal{B} \triangleq \{\mathbf{x}\in\mathcal{A}^{(m)} : \mu^{(m)}(A_{\mathbf{x}}) > \beta^{(m)} - c^{(m)}\}$$

and note that for each i we have $\mathcal{B}_i \subset \mathcal{B}$. It follows that

$$p_i^{(m)} \leq \sum_{\mathbf{x}\in\mathcal{B}}\hat{\pi}^{(m)}(\mathbf{x}).$$

Now, suppose that $\beta^{(m)}$ scales as $(1+\epsilon)\rho m$ for some $\epsilon > 0$. Then, from Theorem 1, the denominator in (9), namely $\sum_{\mathbf{x}\in\mathcal{A}^{(m)}}\pi^{(m)}(\mathbf{x})$, tends to 1 as $m\to\infty$. Thus, the impact of truncation by network capacity becomes negligible as the system becomes large. More formally, the theorem below follows by noting that $c^{(m)}$ is $o(m)$.

Theorem 2 *Consider a sequence of content delivery systems as in Theorem 1. Further suppose that the servers are connected to the users via a shared network link of bandwidth* $\beta^{(m)} = (1+\epsilon)\rho m$ *for some* $\epsilon > 0$*. Suppose if upon arrival of a job the aggregate service capacity* $\mu^{(m)}(A_{\mathbf{X}^{(m)}})$ *exceeds* $\beta^{(m)}$ *then the job is blocked.*

For each m*, the system is stationary. Under stationary distribution, let* $\hat{Y}_s^{(m)}$ *represent the instantaneous activity of server* s*. Let* $p_i^{(m)}$ *be the probability that a class* i *job is blocked. Then, the following statements hold:*

(a) $\lim_{m\to\infty} \sup_i p_i^{(m)} \to 0.$

(b) $\lim_{m\to\infty} \frac{\sum_{l=1}^{m} \hat{Y}_{s_l}^{(m)}}{m} = E\left[\hat{Y}_{s_1}^{(m)}\right]$ *in probability.*

5 Conclusions

The aim of in this paper was to further our understanding of server dynamics in large scale systems which explicitly leverage resource pools, and the related resource allocation problems. There is a scope of improving these results on at least two fronts:

1. We use a weak law of large numbers for a triangular array of exchangeable variables to show concentration in server activity. The concentration may be made more precise by using central limit theorems for such random variables (see Corollary 20.10 in [1]).
2. For content delivery systems, we studied the impact of a shared network link for a case where the link bandwidth is $O(m)$ greater than the average traffic demand. One may strengthen this result by considering cases where the link bandwidth is closer to or lower than the average traffic demand.

Acknowledgements Authors would like to thank François Baccelli and Sanjay Shakkottai at The University of Texas at Austin, and Xiaoqing Zhu at Cisco Systems for helpful discussions which motivated this work. Virag Shah would also like to thank Anurag Kumar at Indian Institute of Science for providing him with the first exposure to mean field models.

References

1. Aldous, D.J.: Exchangeability and related topics. Springer (1985)
2. Barroso, L., Hölzle, U.: The datacenter as a computer: An introduction to the design of warehouse-scale machines. Synthesis Lectures on Computer Architecture **4**(1), 1–108 (2009)
3. Bonald, T., Massoulié, L., Proutière, A., Virtamo, J.: A queueing analysis of max-min fairness, proportional fairness and balanced fairness. Queueing Systems **53**, 65–84 (2006)
4. Bonald, T., Proutière, A.: Insensitive bandwidth sharing in data networks. Queueing Systems **44**, 69–100 (2003)
5. Bonald, T., Proutière, A.: On performance bounds for the integration of elastic and adaptive streaming flows. In: Proceedings of ACM Sigmetrics, pp. 235–245 (2004)
6. Bonald, T., Proutière, A., Roberts, J., Virtamo, J.: Computational aspects of balanced fairness. In: Proceedings of ITC (2003)

7. Bramson, M., Lu, Y., Prabhakar, B.: Asymptotic independence of queues under randomized load balancing. Queueing Systems **71**(3), 247–292 (2012)
8. Edmonds, J.: Submodular functions, matroids, and certain polyhedra. In: Proceedings of Calgary International Conference on Combinatorial Structures and Applications, pp. 69–87 (1969)
9. Kallenberg, O.: Canonical representations and convergence criteria for processes with interchangeable increments. Zeitschrift fr Wahrscheinlichkeitstheorie und Verwandte Gebiete **27**(1), 23–36 (1973)
10. Kelly, F.P.: Reversibility and Stochastic Networks. Wiley (1979)
11. Kelly, F.P.: Loss networks. Ann. Appl. Probab. **1**(3), 319–378 (1991)
12. Lin, M., Wierman, A., Andrew, L., Thereska, E.: Dynamic right-sizing for power-proportional data centers. In: Proceedings of IEEE Infocom, pp. 1098–1106 (2011)
13. Massoulié, L.: Structural properties of proportional fairness: Stability and insensitivity. Annals of Applied Probability **17**(3), 809–839 (2007)
14. Mitzenmacher, M.D.: The power of two choices in randomized load balancing. Ph.D. thesis, University of California, Berkeley (1996)
15. Nemhauser, G.L., Wolsey, L.A.: Integer and combinatorial optimization, vol. 18. Wiley (1988)
16. Shah, V., de Veciana, G.: Performance evaluation and asymptotics for content delivery networks. In: IEEE Infocom, pp. 2607–2615 (2014)
17. Shah, V., de Veciana, G.: Impact of fairness and heterogeneity on delays in large-scale content delivery networks. In: ACM Sigmetrics (2015)
18. Vvedenskaya, N.D., Dobrushin, R.L., Karpelevich, F.I.: Queueing system with selection of the shortest of two queues: An asymptotic approach. Problemy Peredachi Informatsii **32**(1), 20–34 (1996)
19. Whitt, W.: Blocking when service is required from several facilities simultaneously. AT&T Technical Journal **64**(8), 1807–1856 (1985)