Bayesian Admission Policies for Cloud Computing Clusters

Ludwig Dierks\textsuperscript{1}, Ian Kash\textsuperscript{2}, and Sven Seuken\textsuperscript{1}
\textsuperscript{1} Department of Informatics, University Zurich
\textsuperscript{2} Microsoft Research
dierks@ifi.uzh.ch, iankash@uic.edu, seuken@ifi.uzh.ch

Abstract
Cloud computing providers must handle heterogeneous customer workloads for resources such as (virtual) CPU or GPU cores. This is particularly challenging if customers, who are already running a job on a cluster, scale their resource usage up and down over time. The provider therefore has to continuously decide whether she can add additional workloads to a given cluster or if doing so would impact existing workloads’ ability to scale. Currently, this is often done using simple threshold policies to reserve large parts of each cluster, which leads to low average utilization of the cluster. In this paper, we propose more sophisticated Bayesian policies for controlling admission to a cluster and demonstrate that they significantly increase cluster utilization. We first introduce the cluster admission problem and formalize it as a constrained Partially Observable Markov Decision Problem (POMDP). We then systematically design heuristic Bayesian admission policies that estimate moments of each workload’s distribution of future resource usage. Via simulations we show that our Bayesian admission policies lead to a substantial improvement over the simple threshold policy. We then evaluate how much further this can be improved with learned or elicited prior information and how to incentivize users to provide this information.

1 Introduction
Cloud computing is a fast expanding market with high competition where small efficiency gains translate to multi-billion dollar profits.\textsuperscript{1} Nonetheless, most cloud clusters currently run at low average utilization (i.e. only a relatively low fraction of resources is actually used by customers). Some of this is caused by purely technical limitations, such as the need to reserve capacity for node failures or maintenance, outside factors such as fluctuations in overall demand, or inefficiencies in scheduling procedures, especially if virtual machines (VMs) might change size or do not use all of their requested capacity (Yan et al. 2016). Another cause is the nature of many modern workloads: highly connected tasks running on different VMs that should be run on one cluster to minimize latency and bandwidth use (Cortez et al. 2017). In practice, this means that different VMs from one user are bundled together into a deployment of interdependent workload.

In this paper, we pay special attention to the fact that, when the workload of a deployment changes, it can request a scale out in the form of additional VMs or shut some of its active VMs down. To provide a sense of the difficulty of this problem: over time, the number of VMs needed by a specific deployment could vary by a factor 10 or even 100, and a request to scale out should almost always be accepted on the same cluster, as denying them would impair the quality of the service, possibly alienating customers. This means that providers face a difficult problem of deciding to which cluster to assign a deployment, as a deployment which is small today may without warning have a dramatic increase in size that must be accommodated. Providers consequently hold large parts of any cluster as idle reserves to guarantee that only a very low percentage of these requests is ever denied.

1.1 Cluster Admission Control
We reduce the problem of determining to which cluster to assign a new deployment to the problem of determining, for a particular cluster, whether it is safe to admit a deployment, or if doing so would risk running out of capacity if the deployment scales (Cortez et al. 2017). While a lot of research

\textsuperscript{1}https://www.microsoft.com/en-us/Investor/earnings/FY-2018-Q2/press-release-webcast
has been done on scheduling inside the cluster (Schwarzkopf et al. 2013; Verma et al. 2015; Tumanov et al. 2016), the admission decision has not been well studied before. Consequently, cloud providers are still often using simple policies like thresholds on the current utilization of a cluster. These seem reasonable at first glance, as the law of large numbers might suggest that the overall utilization carries most information for large clusters. But as Cortez et al. (Cortez et al. 2013; Verma et al. 2015; Schwarzkopf et al. 2016), the large scale of the problem as well as the highly complicated underlying stochastic processes make finding optimal policies infeasible, even for the underlying Markov Decision Process and with limited look-ahead horizon.

We formalize the cluster admission problem as a constrained Partially Observable Markov Decision Process (POMDP) (Smallwood and Sondik 1973) where each deployment behaves according to some stochastic process and the controlling agent (the cloud provider) tries to maximize the number of active compute cores without exceeding the cluster’s capacity. Since the exact stochastic processes of arriving deployments are not known to the agent, it has to reason about the observed behavior. The large scale of the problem as well as the highly complicated underlying stochastic processes make finding optimal policies infeasible, even for the underlying Markov Decision Process and with limited look-ahead horizon.

1.2 Overview of Contributions

We present tailored heuristics to solve the cluster admission control problem. For this, we first present a succinct mathematical formulation of the cluster admission problem as faced by cloud providers before defining the optimal policy as the solution to a constrained POMDP. We then give a statistical model for the typical population of deployments in cluster and for how each deployment changes over time. We base this on an extensive analysis and fitting of a data trace from a real-world cloud computing center (Microsoft Azure internal jobs (Cortez et al. 2017)).

Next, we propose a strategy for constructing heuristic policies via a series of simplifying assumptions. These assumptions reduce the highly branching look-ahead space down to the approximation of a random variable using its moments. We then present the currently used threshold policy that doesn’t take probabilistic information into account as well as two new policies that take successively higher moments into account. Via simulations, we show that our higher moment policies produce a 4.25% improvement over current practices, which can translate to hundreds of millions of dollars of savings for large cloud providers.

Then we examine how the utilization of the cluster can further be improved if more precise prior information about arriving deployments is available. We give a simple framework which captures a notion of the quality of information available, and through additional simulations quantify how the policies benefit from this additional information. Finally, we present a new information elicitation approach, introducing a variance-based pricing rule to elicit labels from users. This rule provides users with the right incentives to (a) label their deployments properly (into high and low variance deployments) and (b) structure their workloads in a way that helps the cluster run more efficiently.

1.3 Related work

While studying which deployments to admit to a cluster, we abstract away the question of exactly which resources they should use. A literature on cluster scheduling addresses this (Schwarzkopf et al. 2013; Verma et al. 2015; Tumanov et al. 2016). In addition, there is work that addresses a different notion of admission control to a cluster, namely how to manage queues for workloads which will ultimately be deployed to that cluster (Delimitrou et al. 2013).

One line of work for cluster scheduling has looked at how multidimensional resources can be fairly divided among deployments. Dominant Resource Fairness (Ghodsi et al. 2011) is an approach that has proved useful in practice (Hindman et al. 2011) and has inspired follow-up work more broadly in the literature on fair division (Parkes et al. 2012; Dolev et al. 2012; Gutman and Nisan 2012; Kash et al. 2014).

There is also a literature that views scheduling through the lens of stochastic online bin packing (Cohen et al. 2016; Song et al. 2014), which deals with the related issues of changing workloads on often overcommitted resources, but at smaller scales.

In this paper, we examine the value of learning from prior deployments. Other work has explored similar opportunities in the context of resource planning and scheduling in analyt-
ics clusters (Jyothi et al. 2016; Rajan et al. 2016).

There is a large literature on market design challenges in the context of the cloud (Kash and Key 2016). Existing work has studied both queueing models where decisions are made online with no consideration of the future (Abhishek et al. 2012; Dierks and Seuken 2016) and reservation models which assume very strong information about the future (Azar et al. 2015; Babaioff et al. 2017). Our work sits in an interesting intermediate position where users may have rough information about the types of their deployments.

Solving POMDPs is a well-studied problem (Smith and Simmons 2012; Russell and Norvig 2016; Roy et al. 2005). Unfortunately, finding an optimal policy is known to be PSPACE-complete even for finite-horizon problems (Padimitriou and Tsitsiklis 1987). Even finding $\epsilon$-optimal policies is $NP$-hard for any fixed $\epsilon$ (Lasena et al. 2001).

In our case, the problem is further exacerbated by the existence of side constraints. Constrained POMDPs are far less well studied than unconstrained POMDPs. General (approximation) strategies proposed in the past include linear programming (Poupart et al. 2015; Walraven and Spaan 2018), point-based value iteration (Kim et al. 2011), a mix of online-look ahead and offline risk evaluation (Undurti and How 2010), and forward search with pruning (Santana et al. 2016). None of these approaches is efficiently applicable when the state space of the underlying MDP is large or, as in our case, partly continuous. While some work has addressed continuous state space POMDPs (Porta et al. 2006; Duff and Barto 2002; Brooks et al. 2006), none of them are directly applicable to a constrained problem of this size.

2 Preliminaries

2.1 The cluster admission problem

We consider a single cluster in a cloud computing center. A cluster consists of $c$ cores that are available to perform work. $c$ is also called the cluster’s capacity. These cores are used by deployments, i.e. interdependent workloads that use one or more cores. The set of deployments currently on the cluster is denoted by $X$, and each deployment $x \in X$ is assigned a number of cores $C^x$. Any core that is assigned to a deployment is called active, while the remainder are inactive. All inactive cores are assumed to be ready to be assigned and become active at any time.$^3$

The exact placement of cores inside the cluster is not taken into account at this level and in consequence we do not model the grouping of cores into VMs. A deployment can request to scale out, i.e. increase its number of active cores. Each such request is for one or more additional cores and must be accepted whenever activating the requested number of cores would not make the cluster run over capacity. Following current practice, scale out requests must be granted entirely or not at all. Deployments may also stop using some of their cores over time and these cores then become inactive. A deployment dies when its number of active cores becomes zero. It can also die spontaneously, instantly making all of its cores inactive. Intuitively this models a decision by a user to shut down the deployment.$^4$

We assume that the processes for the time between scale out and the time until a core becomes inactive are both memoryless. This is common whenever arrival and departure processes are modeled (e.g., in queuing theory), and has been used in previous models of cloud computing (Abhishek et al. 2012; Dierks and Seuken 2016). New deployment requests arrive over time and are accepted or rejected according to an admission policy based on the current state of the cluster and the arriving deployment. The policy has to make sure that the cluster is not forced to reject a higher percentage of scale out requests than is specified by the internal service level agreement (SLA) $\tau$. If a scale out request cannot be accepted because the cluster is already at capacity, one failure for the purpose of meeting the SLA is logged per requested core. An optimal policy therefore maximizes the utilization of the cluster, i.e. the average number of active cores, while making sure the SLA is observed in expectation (i.e., over time).

2.2 POMDP formulation

If the parameters governing the scale out behavior of deployments were known, the optimal policy could be formulated as the solution to a Markov Decision Problem. Since these parameters are generally not known, the cluster can only reason about them based on observed behavior and (optionally) some a priori information about individual deployments. The problem of finding an optimal policy can therefore be formalized as a constrained discrete time for maintenance, are independent of the cluster admission policy. They therefore do not affect the relative efficiency of policies and don’t need to be modeled.

$^3$While this is an abstraction, effects that make inactive cores become unavailable, such as hardware failure or capacity reserved

$^4$We model death as permanent because with no active cores any future request could be assigned to a different cluster.
Partially Observable Markov Decision Process (POMDP) \((\mathcal{S}, A, \mathbb{R}, T, \mathcal{C}, \tau, \Omega, \mathcal{O})\). While deployments can arrive at arbitrary times, it takes time to make the acceptance decision so there is little loss in assuming for the purposes of our POMDP that time is discrete. \(\mathcal{S}\) denotes the space of all possible states the cluster can be in. A state \(s \in \mathcal{S}\) is assumed to contain all information about the cluster’s active deployments \(X(s)\) (including both their current size and scaling process parameters) as well as the deployments that arrived during the current time step. The action set \(\mathcal{A}\) consists of accepting or rejecting the deployments that arrived this time step. The reward function \(\mathbb{R}(s) = \sum_{x \in X(s)} C^x\) is simply the number of active cores in a state \(s\), \(T(s'|s, a)\forall s' s \in \mathcal{S}, \forall a \in \mathcal{A}\) denotes the state transition probabilities and \(C(s, a)\) denotes the expected number of failures that occur with the next state transition from a given state-action pair \((\mathbb{R} \text{ and } C \text{ require a model of deployment evolution, which will be described in the coming sections})\). \(\tau\) denotes the SLA, as described above. \(\Omega\) is the set of possible observations and \(\mathcal{O}\) an observation model.

The provider does not actually observe the full state space. Instead, an observation \(o \in \Omega\) only reveals whether deployments arrived to the cluster or died and the current size \(C^x\) of any active deployment \(x \in X\), while in particular not revealing the parameters of the deployments. Note that this means that our observation model \(\mathcal{O}\) is deterministic, but that many states share the same observation. As is standard, we further denote the clusters’ current knowledge about which state \(s\) it is in via a belief state \(B\), i.e. a probability distribution over all possible states. Whenever the state of the system changes in some time step \(n\) and the cluster obtains a new observation \(o\), the belief state is updated according to the observation model, i.e.

\[
b_{n+1}(s' | b_n, a, o) \propto \mathcal{O}(o | s') \sum_s T(s'|s, a)b_n(s).
\]

We can now formulate the problem of finding an optimal policy.

**Problem 2.1.** An optimal policy \(p^*\) starting in belief state \(b\) can be found by solving

\[
p^* = \arg\max_p \sum_n \int_s f_{n, p, b}(s) \sum_{x \in X(s)} C^x ds \tag{2}
\]

\[
\text{s.t. } \int_s f_{n, p, b}(s) C(s, a_p) ds \leq \tau \forall n > 0 \tag{3}
\]

where \(f_{n, p, b}\) is the probability density function of the distribution over the states of the system for timestep \(n\) given policy \(p\) and starting belief \(n\). An optimal policy therefore maximizes the expected reward, i.e. the expected number of active cores over all future timesteps. The side constraint \((3)\) guarantees that the optimal policy is only chosen from those policies that observe the SLA, given the provider’s belief.

Before moving on to fit a model of deployments, we pause to examine some key features of our model. Note that if the underlying state was fully observable, with \(c\) cores there could be as many as \(c\) deployments each of which requires (as we will see) at least four parameters to describe its size and processes. The parameter space for each stochastic process is continuous, meaning a state \(s \in \mathcal{S}\) consists of more than \(c\) discrete and \(3c\) continuous parameters. Thus, even if we would fully discretize the parameter space, the number of states of our underlying MDP would be exponential in \(c\). This makes solving it a challenge even without the added complexity of the SLA constraint and the partial observability. This means that approaches to solving our POMDP which would require solving versions of the MDP as a subroutine are infeasible in our setting, where clusters contain 10000 or more cores.

### 2.3 Data Trace

To have a better understanding of the scaling behavior of real deployments and find proper representations for the state and belief state, we fitted the behavior of deployments to a real-world data trace. We will later use the fitted processes to power our simulations in Sections 4 and 5. The particular data trace we use was published by Cortez et al. (Cortez et al. 2017).\(^5\) This dataset consists of one month of data of internal Microsoft Azure jobs. It contains 35,576 deployments, but only 29,757 of these, those that arrived during the observed time period, can be used for fitting without getting skewed towards long lived deployments. The rest were culled from the dataset. The 29,757 deployments activated 4,317,961 cores, out of which 4,211,926 became inactive again during the observed month. The exact lifetime of the remaining cores (i.e. the length of time between becoming active and then inactive again) is not known; instead we only have a lower bound on it (i.e. our observation is Type I censored; see for example (NIST 2012)). Thus, for cores where

\(^5\)In contrast to (Cortez et al. 2017) we did not consolidate all deployments a single user runs on a certain day into one. This is because cores that get requested as a new deployment do not need to be accepted on the same cluster.
we only have a lower bound on the lifetime we use the cdf in our likelihood function while for cores whose lifetime is known we use the pdf.

2.4 Fitting on the deployment level

We first fit arrival and departure processes for each individual deployment. In keeping with the Markov assumption, we fit a Poisson distribution to the scale out rate of each deployment, while we fit an exponential distribution to the lifetime of cores for each deployment for which at least one core became inactive during the observed time period. Note that while we model the cluster admission problem as a discrete-time POMDP, the processes are fit in continuous time. This is more general, avoids imprecisions introduced by time discretization and easily allows choosing any discretization after fitting. To fit the size of a scale out, we also used a Poisson distribution (plus 1, as scale outs must have at least one core).

2.5 Fitting on the population level

Once we had distributions for each deployment we turned to fitting prior distributions for the population. The parameters of the processes for each arriving deployment are drawn from these populations. They further inform the starting belief state of the provider over the true parameters of his active deployments, i.e. the provider’s prior for any arriving deployment consists of the knowledge that it was drawn from this distribution. As the data was skewed, positive, and not really heavy tailed, a Gamma Distribution is a natural and very general candidate (containing the Chi-squared, Erlang and Exponential distributions as special cases). We compared it to a few other potential distributions (Weibull, Lognormal, Lomax) and it indeed resulted in the best fit for each process by a large margin. The resulting model and parameters from our fits are shown in Table 1. While the scale out size is fit directly to the samples, scale out rate and core lifetime are highly correlated. The longer a deployment’s cores live, the lower the rate at which new cores arrive, as can be seen in Figure 3. This is interesting, as it means that deployments with long lived cores do not necessarily have more active cores. To account for it, we normalized all scale out rates by the respective core lifetimes.

To visualize the fitted distributions, Figure 1 shows the cdf of the Gamma Distribution for the lifetime parameter, overlaid over the normalized cumulative histogram of the fitted rates of the sample deployments. Figure 2 shows the cdf over the lifetime of a core from a random deployment in our model over the normalized cumulative histogram of one core, drawn randomly from each deployment that started in the first seven days. It only shows cores that are started early, as the realized lifetime of later cores is evermore skewed by the time left until the end of the observed month. Visually, our prior produces an excellent fit to the data.

Figure 4 shows the cdf for the normalized scale out rates over the relevant cumulative histogram. The actual scale out rate of a sampled deployment is now simply the normalized scale out rate multiplied by the average core lifetime. Figure 5 shows the fitted cdf for scale out size parameters over its cumulative histogram. Figure 6 shows the probability mass function of drawing a random scale out from a randomly drawn deployment in our model over the histogram of the size of one scale out event drawn randomly from each deployment in the data set. This shows that the mix of scale out sizes in the simulated population based on our priors is roughly equal to the data trace.

Deployment Shutdown. While most deployments in the dataset die because they have zero active cores, 5,980 of the 22,241 deployments that both arrive and die during the observed period seem to get actively shut down. By this we mean that they had at least 3 VMs that all shut down simultaneously. This would be highly unlikely if deployments only

---

6As the Poisson distribution is single-parameter and its variance cannot be set independent of the average size, this is not a particularly good fit for users with large but consistent scale out sizes. However, its simplicity avoids overfitting on the often low number of samples per deployment and it results in a good fit on the population level.

7Note that this cdf takes the form of a Lomax distribution as the convolution of a Gamma and an Exponential distribution.

8Note that Azure only offers VMs whose core counts are a power of two.
Table 1: Fitted processes

| Deployment processes | scale out process | scale out size distribution | core lifetime distribution | deployment lifetime distribution |
|----------------------|-------------------|-----------------------------|---------------------------|-------------------------------|
| Distribution         | Poisson($\Lambda M$) | Poisson($\Sigma$) | Exponential($M$) | Exponential($\Delta M$) |
| Deployment parameters| normalized scale out rate $\Lambda$ | average scale out size $\Sigma$ | core lifetime rate $M$ | deployment lifetime rate per core lifetime $\Delta$ |
| Prior                | Gamma(0.3494, 0.1543) | Gamma(0.2705, 0.0571) | Gamma(0.3190, 0.5932) | 0.0337 |

Figure 3: Scale out rate as a function of average core lifetime

Figure 4: Distribution of scale out rate parameters

Figure 5: Distribution of scale out size parameters

Figure 6: Distribution of scale out sizes

die when cores or VMs become inactive independently. To capture such behavior we fit an exponential distribution over the number of expected core lifetime deployments lived. The maximal lifetime of deployments that did not get shut down was assumed to be censored to their realized lifetime.

3 Admission Control Policies

In this section, we present the simple cluster admission policy currently used in practice, as well as our new, more sophisticated Bayesian policies.

Even with limited look-ahead horizons, optimal policies cannot feasibly be calculated for the POMDP. There is no simple closed form for the state transition probabilities and even if there was, the branching factor of the POMDP is too large. This problem persists even if we would only try to solve the underlying MDP. We therefore present heuristic policies tailored to the cluster admission problem. Our policies are based on a number of carefully/systematically chosen assumptions:

Disregard future arrivals Our policy does not reject deployments simply because there is a chance that better behaved deployments arrive in the future. The optimal policy, under this assumption, accepts a new deployment whenever doing so does not violate the side constraint. This allows our policy to decide acceptance based only on the failure probability caused by the current population of the cluster and assuming no more deployments arrive in the future. In the cloud domain, this assumption is actually desirable, as even customers with high demand variability must be served by some cluster in the data center.

Relax capacity constraint. With disregarding future arrivals, the cluster’s future state only depends on how the sizes of the current active deployments change. Despite this, the probabilities are complex because if one deployment helps fill the cluster then further scale out requests even by other deployments are denied, introducing correlations between the size of deployments. Instead, the policy assumes the cluster can run an infinite number of cores, which allows the evolution of each deployment to be independent. Intuitively this is reasonable because the cluster being full should be rare if the constraint on the policy is being met. Note that we only make this simplifying assumption for the prediction of state transitions; scale outs that would fail in reality are still logged as failures by the policy.

With these first two assumptions, we can now describe the future evolution of the cluster using the following random variables (which have a superscript to indicate deployment $x$ being referenced which we omit for brevity).

- $C$ is the variable denoting the number of active cores at
• $Y_i$ is the random variable denoting the number of scale outs that occur between timestep $i - 1$ and timestep $i$, assuming the deployment has not died.
• $S_{i,l}$ is the size of the $l$'th scale out request that would have occurred at timestep $i$.
• $Z_{n,i,k}$ is the binary random variable denoting whether the $k$'th core activated between timesteps $i - 1$ and $i$ would still be active in timestep $n$, assuming at least $k$ cores were activated and the deployment has not died. If $i = 0$, this refers to the set of cores already active at timestep 0.
• $D_i$ is the random variable which is 1 if $x$ would not have died due to a lack of active cores before timestep $i$. It can be defined recursively as
\[
D_i = D_{i-1}(1 - \prod_{k=1}^{C} (1 - Z_{i,0,k})) \prod_{j=1}^{i-1} \sum_{l=0}^{n} S_{j,l} \prod_{k=1}^{C} (1 - Z_{i,j,k}))
\]
• $B_n$ is the random variable denoting the number of cores that were active at timestep 0 and are still active in timestep $n$, which can be calculated as
\[
B_n = \sum_{k=1}^{C} Z_{n,0,k}.
\]
• $Q_n$ is the random variable denoting the number of cores activated between timestep 0 and timestep $n - 1$ that are still active assuming no service termination, i.e.
\[
Q_n = \sum_{i=1}^{n-1} \sum_{k=1}^{C} S_{i,k} Z_{n,i,k}.
\]
• $\Omega_i$ is the random variable which is 1 if the maximum lifetime of the deployment is at least $i$ and 0 otherwise.
• Finally, $L_n$ is the random variable denoting the number of active cores in timestep $n$ which can be calculated as
\[
L_n = \Omega_n D_n (Q_n + B_n).
\]

**Proposition 1.** Under the three assumptions (disregard future arrivals, relax capacity constraint, and assume at most one failure occurs per timestep), the side constraint (3) for any timestep $n$ becomes
\[
\int_s f_{n,p,b}(s) C(s, a_p) \, ds = \Pr(\sum_{x} L_{n+1}^{x} > c) \leq \tau. \quad (10)
\]

An optimal policy that operates under these three assumptions accepts an arriving deployment whenever doing so does not violate Inequality (10).

The proof follows directly from the three assumptions and is omitted due to space constraints.

**Approximate failure probability using (approximate) moments** While the last three assumptions greatly simplified the problem, the processes at hand leave us without an analytical form representation of Inequality (10). Instead, we utilize the (approximate) moments of $L_n$ to construct simple policies. In the following, we present three such policies that make use of successively higher moments.

### 3.1 Zeroth Moment Policy (Baseline)

The baseline for admission control policies that is widely used in practice is a myopic policy that simply compares the current number of active cores to a threshold. This policy does not require any information about the set of deployments besides the total number of active cores. We also call it a Zeroth Moment Policy, because it takes no information about the future into account. The limited amount of information it uses means that it has to be very conservative in how many deployments to accept, since it does not know how often or fast they will scale out.

**Definition 1 (Zeroth Moment policy).** With a zeroth moment policy with threshold $t$, a newly arriving deployment is accepted if after accepting the deployment there would be less than $t$ cores active in the cluster.

### 3.2 First Moment Policy

First moment policies try to approximate (10) by utilizing the first moments, i.e. the mean, of the deployment processes. In the spirit of Markov’s inequality, we propose to reject an arriving deployment when the expected utilization lies above some chosen threshold and otherwise accept it.

**Definition 2 (First Moment Policy).** With a first moment policy with threshold $t$, a newly arriving deployment is ac-
cepted whenever after accepting the deployment the expected number of active cores would be less than \( t \) in all future timesteps, i.e.

\[
\sum_{x \in X} E[L_n^x] \leq t \quad \forall n
\]  

(11)

Note that, unless the exact parameters of a deployment are known, \( Y^x_t, S^x_t \) and \( Z_{n,i,k} \) for a single deployment \( x \) are highly correlated. While the expectations of \( Q_n \) and \( B_n \) can easily be expressed in the expectations of their component random variables, \( D_n \) is a compound of other processes and very hard to calculate for most distributions. In practice, it therefore is usually preferable to only approximate \( D_n \). Similarly, ignoring the correlation between the \( Q_n \), \( D_n \) and \( B_n \) greatly simplifies policy evaluation and does not overly impact the prediction quality. Thus, our policy uses the following mix of exact and approximate moments.

**Proposition 2.** It holds:
\[
E[L_n] \approx E[\Omega_n]E[D_n](E[Q_n] + E[B_n])
\]  

(12)

\[
E[Q_n] = \sum_{i=1}^{n-1} E[Y_i|\lambda, \mu]E[S_{1,1}|\sigma]E[Z_{n,i,1}|\mu]
\]  

(13)

\[
E[D_i] \approx E[D_{i-1}](1 - (1 - E[Z_{i,0,1}])^{(C)})
\]  

(14)

\[
E[D_{1}] = \prod_{j=1}^{i-1}(1 - (1 - E[Z_{i,j,1}])^{E[Y_{i,1}|E[S_{1,1}])})
\]  

(15)

\[
E[B_n] = CE[Z_{n,i,k}]
\]  

(16)

The proof and the expressions for the specific distributions fitted to the data trace can be found in the appendix.

### 3.3 Second Moment Policy

First moment policies still fail to take into account much of the structure of deployments. In a sense they always have to take the worst possible population mix into account and run the risk of accepting deployments with low expected size but high variance when close to the threshold. One way around this is to also take the second moment, i.e. the variance of \( L_n \), into account. To address this, we propose to use Cantelli’s inequality, a single-tailed generalization of Chebyshev’s inequality, to approximate Inequality (10).

**Proposition 3 (Cantelli’s Inequality).** For any real-valued random variable \( L \) and any \( \epsilon \geq 0 \) it holds

\[
Pr(L - E[L] \geq \epsilon) \leq \frac{Var[L]}{Var[L] + \epsilon^2}
\]  

(18)

If we now set \( \epsilon = (c - \sum_{x \in X} E[L_n^x]) \), we can bound the probability of running over capacity. While the bound given by the inequality is not tight enough to simply set it to \( \tau \), it can be used to bound the first two moments in a systematic way.

**Definition 3 (Second Moment Policy (Cantelli Policy)).** With a Cantelli policy with threshold \( \rho \), a newly arriving deployment is accepted if the estimated probability of running over capacity in all further timesteps is less than \( \rho \), i.e.

\[
\sum_{x \in X} Var[L_n^x] + (c - \sum_{x \in X} E[L_n^x])^2 \leq \rho \quad \forall n
\]  

(19)

The variance is more complex to calculate than the expectation. Our general approximation result, again ignoring some of the covariance, is given by the next proposition.

**Proposition 4.** It holds:

\[
V[L_n] = E[\Omega_n]V[D_nQ_n] + E[\Omega_n](1 - E[\Omega_n])E[D_nQ_n]^2
\]  

(20)

\[
V[D_nQ_n] \approx E[D_n]V[Q_n] + E[Q_n]^2E[D_n](1 - E[D_n])
\]  

(21)

\[
V[Q_n] = E[|Q_n\lambda, \sigma, \mu|] + VE[|Q_n\lambda, \sigma, \mu|]
\]  

(22)

\[
C V[Q_n, \lambda, \sigma, \mu] = \sum_{i=1}^{n-1} ((V[Y_i|\lambda, \mu]E[S_{i,1}|\sigma])^2 + E[Y_i|\lambda, \mu]E[S_{i,1}|\sigma]E[Z_{n,i,1}|\mu])
\]  

(23)

\[
V[D_n] \approx E[D_n](1 - E[D_n])
\]  

(24)

\[
V[B_n] \approx CV[Z_{n,i,k}]
\]  

(25)

The proof and the expressions for the specific distributions fitted to the data trace can be found in the appendix.

**Computational overhead.** The computational overhead of the second moment policy depends on the number of future timesteps it evaluates. Fortunately, each single rule application is fast. Updating the estimate for the second moment policy depends on the number of functions fitted to the data trace can be found in the appendix. A cloud computing center consisting of clusters of capacity \( c \) with an arrival rate of \( L \) new deployment requests per hour therefore has a computational overhead of at most \( O(Lcn^2) \) each hour, parallelizable.
into jobs of size $O(n^2)$. This number can further be significantly reduced in practice by employing smart caching and by updating the estimates only whenever parameters change significantly.\footnote{If further ML is (optionally) employed to obtain an individual prior for arriving deployments (as discussed in Section 5), that computation time would need to be added and depends on the algorithm in question.} This means that even relatively large look-ahead horizons $n$ can easily be implemented in practice.

## 4 Empirical Evaluation

In this section, we evaluate the performance of our admission policies using the fitted model we have presented in Section 2.5. This provides a performance comparison between our policies and the industry baseline of the zeroth moment policy.\footnote{Note that our dataset only covers one month. This is too short to fully evaluate cluster admission policies because many effects only show up after months of usage. Instead, we use our behavioral model, fitted to the data we do have, to simulate longer time periods (1 year, in our simulations). We defer evaluations against real deployments to future work once more data is available.} We simulate clusters with capacity $c = 10,000$ for a 1-year period with all three policies. We determine the optimal threshold for each policy via binary search, subject to meeting an SLA of $0.01\%$. An average of 6 new deployments per hour arrives according to a Poisson process. The parameters of each arriving deployment are drawn from the fitted distributions presented in Table 1.

To simulate a full year with a reasonable number of core-hours, the first and second moment policies both only look ahead 2 hours, discretized in 12 timesteps. This means that the policies assume that new scale out requests only arrive every 10 minutes. Note that such a short look ahead would not be required in practice, as calculating the same precision of a single deployment for one month only takes 5 core seconds and only needs to be done whenever a deployment’s learned information changes notably. Each simulation was repeated 20 times. Each simulation was run on a single core of an Intel E5-2680 v2 2.80GHz processor. The combined runtime of all simulations was about 1,000 core hours.

The results are summarized in Table 2. The zeroth moment policy obtains its best result with a threshold of $t = 8564$, i.e. new deployments are accepted whenever less than 8563 would be active in case of acceptance. This results in an average utilization of 87.45 over the lifetime of the cluster. Note that this is higher than would be obtained in reality, as we do not model buffers for hardware failures, etc. The first moment policy with threshold $t = 8794$ increases utilization by 1.93 to 89.38, a relative increase by 2.21\%. This means that the same number of deployments can be served by roughly 2\% less hardware. While this might not seem much at a first glance, reducing hardware requirements by even a single percent saves hundredth of millions of dollars over the hardware lifetime, which go directly into gross profits instead. This is an invaluable competitive edge. The second moment policy with threshold $\rho = 0.00039$ achieves a utilization of 91.1, which constitutes a relative increase of 4.25\% over the zeroth moment policy and of 2\% over the first moment policy. This shows that taking the second moment into account greatly improves the usefulness of learning parameters during runtime.

### 5 Learned or Elicited Prior Information

So far, our observation model assumed that the cluster does not have any information about arriving deployments, except for their initial number of cores. The acceptance decision must therefore primarily depend on the state of the deployments that are already in the cluster. Intuitively, policies could more precisely control whether accepting a deployment would risk violating the SLA if they had more information about its future behavior. One way to obtain such information would be to use machine learning (ML) based on features of the arriving deployment and past deployment patterns of the submitting user (Cortez et al. 2017). While evaluating particular ML algorithms is beyond the scope of this paper, we evaluate the effect different levels of available information have. To do this we need to parametrize the level of knowledge. For this we assume that the cluster simply gets passed some number of observations from the
true distribution of any arriving deployment.\textsuperscript{11}

5.1 Empirical Evaluation of Prior Information

We simulated the first and second moment policies with varying numbers of prior observations. The short look-ahead of 2 hours now poses a problem because long lived, slow scaling deployments are estimated to be very small in the look-ahead period, yet will in fact grow large over time.\textsuperscript{12} Therefore we implemented the following heuristic: any arriving deployment is assumed to have $\Gamma \Sigma M^2$ active cores, the expected average size it would be if being allowed to stay in the cluster forever without dying.\textsuperscript{13} If accepted, a deployment still only arrives with the drawn arrival size equal to one scale out. We run simulations with 5 different levels of information (0,1,5,20,50 observations), with the same setup as in Section 4 and 20 runs each. These simulations took 3300 core-hours. The results are shown in Figure 7. We see that having prior knowledge equivalent to a single sample would improve utilization significantly. Increases the amount of information does then further reduce the amount of cores that stand idle. With the equivalent of 50 observations, utilization is increased by up to 9% over the baseline. This means the provider requires about 9% less hardware to run the same number of deployments, greatly reducing costs and increasing profits. Further, at any level of information the second moment policy tends to outperform the first.

5.2 Elicited Prior

While using machine learning to predict the parameters of deployments is powerful, users typically do not submit deployments with arbitrary parameters. Instead, they may have a small number of different types of deployments. While sophisticated machine learning may be able to learn this, it may be better to directly elicit some information from the users. However, asking for estimates of parameters for a given deployment is problematic, as it either shifts risk to the consumer or enables manipulation. This leads to the idea of getting users to categorize their deployments into different types, with each type consisting of roughly similar deployments. Learning priors for each individual type could then result in more precise priors and higher efficiency.

Typically, users are charged a fixed payment per hour for each core their deployment uses. To this, we now add a small additional charge based on the variance of the estimate for the deployment’s scaling process and allow users to label the type of their deployments, resulting in an hourly payment rule of the form

$$\pi(x) = \kappa_1 C^x + \kappa_2 Var(x), \quad (31)$$

where $\kappa_1$ and $\kappa_2$ are price constants and $Var(x)$ is an estimate of the variance of the deployment. A payment rule of this form incentivizes users to assign similar labels to similar deployments to minimize the estimated variance.

To see this, consider a user who has two types of deployment, $x$ and $y$, with true variances $Var(x)$ and $Var(y)$. He could now simply submit the deployments under a single label. For the provider, this means that each submitted deployment is of either type with a certain probability, which increases the variance of his prediction. But if the user would label his deployments instead, the provider would know for each arriving deployment which type it is drawn from, reducing variation and therefore the need to reserve capacity. The following proposition, which is immediate from the law of total variance, then shows that, at least in the long run, labeling his deployments also reduces a user’s payments.

**Proposition 5.** Let $z$ be the mixture that results from submitting one of two types of deployments $x$, $y$ chosen by a....
Bernoulli random variable $\alpha \sim \text{Bernoulli}(p_\alpha)$, i.e. such that $z$ is of type $x$ with probability $p_\alpha$ and of type $y$ with probability $1 - p_\alpha$. Then it holds

$$p_\alpha \text{Var}(x) + (1 - p_\alpha) \text{Var}(y) \leq \text{Var}(z) \quad (32)$$

Proof. Since $z$ trivially has finite variance, the law of total variance states:

$$\text{Var}(z) = E[\text{Var}(z|\alpha)] + \text{Var}(E[z|\alpha]) \quad (33)$$

$$\geq E[\text{Var}(z|\alpha)] \quad (34)$$

$$= p_\alpha \text{Var}(x) + (1 - p_\alpha) \text{Var}(y) \quad (35)$$

Proposition 5 shows that the user would be better off, on average, by splitting the mixture and submitting the deployments under separate labels. Note that this abstracts away issues of learning and non-stationary strategic behavior; but for reasonable learning procedures we expect a consistent labeling to lead to lower variance than a mixture while learning. Further, this approach not only gives the user correct incentives to reveal the desired information, but actually incentivizes him to improve the performance of the system. In particular, another way he can lower his payment under this scheme (outside the scope of our model) is to design his deployments in such a way that they have lower variance in their resource use. Since more predictable deployments would allow the policy to maintain a smaller buffer, this provides an additional benefit to the efficiency of the system.

How much any given user could ultimately save by labeling his deployments mostly depends how different his deployment types are and on how high the provider sets the charge for variance. A user whose deployments are very uniform will not save much, while a user with some deployments which never scale and some that scale a lot can potentially save quite a bit. Note that how much the provider should charge is not immediately clear. While he would want to set a high price to put a strong incentive on users, he also has to keep his competition from other providers in mind. At what point the loss of market share outweighs the gain in efficiency is an intriguing problem we leave for future work.

6 Conclusion

We have studied the problem of cluster admission control for cloud computing. The optimal policy would be given as the solution to a very large constrained POMDP, which is infeasible to solve. In practice, simple threshold policies are therefore used for this problem, while we propose carefully designed heuristic policies. Our simulations, with parameters fit to traces from Microsoft Azure, show the potential gains based on our policies. Our results demonstrate that utilization can be increased by up to 4.25% just from learning about deployments while they are active in the cluster. This can be increased to a 9% gain if better prior information about arriving deployments is available, for example through learning or elicitation techniques. While these gains are relatively small in percentage terms, they are economically significant. At cloud scale, these savings translate to hundreds of millions of dollars over the course of a hardware lifetime, and any dollar saved directly translates to a gross profit increase for the cluster provider.

Our overall approach is fundamentally about managing the tail risks of a stochastic process. In our case, these are the rare events where the cluster runs out of capacity. Thus, our approach may also be of interest in other domains where the management of tail risks is important, for example in finance.
References

Vineet Abhishek, Ian A. Kash, and Peter Key. Fixed and market pricing for cloud services. In *7th Workshop on the Economics of Networks, Systems, and Computation (NetEcon)*, pages 157–162, 2012.

Yossi Azar, Inna Kalp-Shaltiel, Brendan Lucier, Ishai Menache, Joseph Naor, and Jonathan Yaniv. Truthful online scheduling with commitments. In EC15, 2015.

Moshe Babaioff, Yishay Mansour, Noam Nisan, Gali Nof, Carlo Curino, Nar Ganapathy, Ishai Menache, Omer Reingold, Moshe Tennenholtz, and Erez Timnat. Era: a framework for economic resource allocation for the cloud. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 635–642. International World Wide Web Conferences Steering Committee, 2017.

Alex Brooks, Alexei Makarenko, Stefan Williams, and Hugh Durrant-Whyte. Parametric pomdps for planning in continuous state spaces. *Robotics and Autonomous Systems*, 54(11):887–897, 2006.

Maxime Cohen, Philipp Keller, Vahab Mirrokni, and Morteza Zadimoghaddam. Overcommitment in cloud services-bin packing with chance constraints. 2016.

Eli Cortez, Anand Bonde, Alexandre Muzio, Mark Russinovich, Marcus Fontoura, and Ricardo Bianchini. Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms. In *Proceedings of the 26th Symposium on Operating Systems Principles*, SOSP ’17, pages 153–167, New York, NY, USA, 2017. ACM.

Christina Delimitrou, Nick Bambos, and Christos Kozyrakis. Qos-aware admission control in heterogeneous datacenters. In ICAC, pages 291–296, 2013.

Ludwig Dierks and Sven Seuken. Cloud pricing: the spot market strikes back. In *The Workshop on Economics of Cloud Computing*, 2016.

Danny Dolev, Dror G Feitelson, Joseph Y Halpern, Raz Kupferman, and Nathan Linial. No justified complaints: On fair sharing of multiple resources. In *proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 68–75. ACM, 2012.

Michael O’Gordon Duff and Andrew Barto. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, University of Massachusetts at Amherst, 2002.

Ali Ghodsi, Matei Zaharia, Benjamin Hindman, Andy Konwinski, Scott Shenker, and Ion Stoica. Dominant resource fairness: Fair allocation of multiple resource types. In *USENIX Symposium on Networked Systems Design and Implementation*, March 2011.

Avital Gutman and Noam Nisan. Fair allocation without trade. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 719–728. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

Benjamin Hindman, Andy Konwinski, Matei Zaharia, Ali Ghodsi, Anthony D Joseph, Randy H Katz, Scott Shenker, and Ion Stoica. Mesos: A platform for fine-grained resource sharing in the data center. In *NSDI*, volume 11, pages 22–22, 2011.

Sangeetha Abdu Jyothi, Carlo Curino, Ishai Menache, Shravan Matthur Narayananmorthy, Alexey Tumanov, Jonathan Yaniv, Ruslan Mavlyutov, Inigo Goiri, Subru Krishnan, Jarnardhan Kulkarni, et al. Morphues: Towards automated slos for enterprise clusters. In *OSDI*, pages 117–134, 2016.

Ian A. Kash and Peter B. Key. Pricing the cloud. *IEEE Internet Computing*, 2016. In Press.

Ian A. Kash, Ariel D. Procaccia, and Nisarg Shah. No agent left behind: Dynamic fair division of multiple resources. *Journal of Artificial Intelligence Research*, 51:579–603, 2014.

Dongho Kim, Jaesong Lee, Kee-Eung Kim, and Pascal Poupart. Point-based value iteration for constrained pomdps. In *IJCAI*, pages 1968–1974, 2011.

Christopher Lusena, Judy Goldsmith, and Martin Mundhenk. Nonapproximability results for partially observable markov decision processes. *Journal of artificial intelligence research*, 14:83–103, 2001.

D. C. Parkes, A. D. Procaccia, and N. Shah. Beyond dominant resource fairness: Extensions, limitations, and indivisibilities. In *ACM Conference on Electronic Commerce (EC)*, June 2012.

NIST. Nist/sematech e-handbook of statistical methods. [http://www.itl.nist.gov/div898/handbook/apr/section4/apr412.htm](http://www.itl.nist.gov/div898/handbook/apr/section4/apr412.htm), 2012.

Christos Papadimitriou and John N. Tsitsiklis. The complexity of markov decision processes. *Math. Oper. Res.*, 12(3):441–450, August 1987.

References
Josep M Porta, Nikos Vlassis, Matthijs TJ Spaan, and Pascal Poupart. Point-based value iteration for continuous pomdps. *Journal of Machine Learning Research*, 7(Nov):2329–2367, 2006.

Pascal Poupart, Aarti Malhotra, Pei Pei, Kee-Eung Kim, Bongseok Goh, and Michael Bowling. Approximate linear programming for constrained partially observable markov decision processes. In *AAAI*, volume 1, pages 3342–3348, 2015.

Kaushik Rajan, Dharmesh Kakadia, Carlo Curino, and Subru Krishnan. Perforator: eloquent performance models for resource optimization. In *Proceedings of the Seventh ACM Symposium on Cloud Computing*, pages 415–427. ACM, 2016.

Nicholas Roy, Geoffrey Gordon, and Sebastian Thrun. Finding approximate pomdp solutions through belief compression. *Journal of artificial intelligence research*, 23:1–40, 2005.

Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited, 2016.

Pedro Santana, Sylvie Thiébaux, and Brian Williams. Rao*: an algorithm for chance constrained pomdps. In *Proc. AAAI Conference on Artificial Intelligence*, 2016.

Malte Schwarzkopf, Andy Konwinski, Michael Abd-El-Malek, and John Wilkes. Omega: flexible, scalable schedulers for large compute clusters. In *Proceedings of the 8th ACM European Conference on Computer Systems*, pages 351–364. ACM, 2013.

Richard D Smallwood and Edward J Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations research*, 21(5):1071–1088, 1973.

Trey Smith and Reid Simmons. Point-based pomdp algorithms: Improved analysis and implementation. *arXiv preprint arXiv:1207.1412*, 2012.

Weijia Song, Zhen Xiao, Qi Chen, and Haipeng Luo. Adaptive resource provisioning for the cloud using online bin packing. *IEEE Transactions on Computers*, 63(11):2647–2660, 2014.

Alexey Tumanov, Timothy Zhu, Jun Woo Park, Michael A Kozuch, Mor Harchol-Balter, and Gregory R Ganger. Tetrisched: global rescheduling with adaptive plan-ahead in dynamic heterogeneous clusters. In *Proceedings of the Eleventh European Conference on Computer Systems*, page 35. ACM, 2016.

Aditya Undurti and Jonathan P How. An online algorithm for constrained pomdps. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 3966–3973. IEEE, 2010.

Abhishek Verma, Luis Pedrosa, Madhukar Korupolu, David Oppenheimer, Eric Tune, and John Wilkes. Large-scale cluster management at google with borg. In *Proceedings of the Tenth European Conference on Computer Systems*, page 18. ACM, 2015.

Erwin Walraven and Matthijs TJ Spaan. Column generation algorithms for constrained pomdps. *Journal of Artificial Intelligence Research*, 62:489–533, 2018.

Ying Yan, Yanjie Gao, Yang Chen, Zhongxin Guo, Bole Chen, and Thomas Moscibroda. Tr-spark: Transient computing for big data analytics. In *SoCC*, 2016.
Appendix

Proposition 6. It holds:

\[
E[L_n] \approx E[\Omega_n]E[D_n](E[Q_n] + E[B_n])
\]

(36)

\[
E[Q_n] = \sum_{i=1}^{n-1} E[E[Y_1|\lambda, \mu]E[S_{i,1}][\sigma]E[Z_{n,i,1}|\mu]]
\]

(37)

\[
E[D_i] \approx E[D_{i-1}](1 - E[Z_{i,0,1}])^{(C)}
\]

\[
E[D_i] \approx E[D_i - 1](1 - (1 - E[Z_{i,0,1}])^{(C)})
\]

Pi \sum_{j=1}^{i-1} (1 - E[Z_{i,j,1}])^{(C)}(38)

(39)

\[
E[D_1] = (1 - (1 - E[Z_{i,0,1}])^{(C)})
\]

(40)

\[
E[B_n] = CE[Z_{n,i,k}]
\]

(41)

Proof. • If \(E[\Omega_n], E[D_n], E[Q_n] \) and \(E[B_n]\) would be uncorrelated, it would hold for the expectation of \(L_n\):

\[
E[L_n] = E[\Omega_n]E[D_n](E[Q_n] + E[B_n])
\]

(42)

(43)

\[
E[L_n] \approx E[\Omega_n]E[D_n](E[Q_n] + E[B_n])
\]

(44)

As the correlation is limited, especially with more information,

\[
E[L_n] \approx E[\Omega_n]E[D_n](E[Q_n] + E[B_n])
\]

(45)

(46)

\[
E[L_n] \approx E[\Omega_n]E[D_n](E[Q_n] + E[B_n])
\]

(47)

is a reasonable approximation in practice.

• \(Q_n\): For the expectation it holds:

\[
E[Q_n] = E[\sum_{i=1}^{n-1} \sum_{k=1}^{Y_i} Z_{n,i,k}]
\]

(48)

\[
= \sum_{i=1}^{n-1} E[\sum_{k=1}^{Y_i} Z_{n,i,k}]
\]

(49)

\[
= \sum_{i=1}^{n-1} E[E[\sum_{k=1}^{Y_i} Z_{n,i,k} | \lambda, \sigma, \mu]]
\]

(50)

\[
E[\sum_{k=1}^{Y_i} Z_{n,i,k} | \lambda, \sigma, \mu] = E\left[\sum_{l=0}^{Y_i} S_{i,l} | \lambda, \sigma, \mu\right]E[Z_{n,i,1} | \lambda, \sigma, \mu]
\]

(51)

\[
= E[\sum_{l=0}^{Y_i} S_{i,l} | \lambda, \sigma, \mu]E[S_{i,1} | \lambda, \sigma, \mu]
\]

(52)

\[
= E[Y_1 | \lambda, \mu]E[S_{i,1} | \sigma] \sum_{i=1}^{n-1} E[Z_{n,i,1} | \mu]
\]

(53)
• $D_n$: If $Y_i, S_{i,l}, Z_{i,j,k}, D_i$ for all $i, j, l, k$ would be uncorrelated, it would hold

\[
E[D_i] = E[D_{i-1}(1 - \Pi_{j=0}^{i-1} \sum_{k=0}^{Y_i} S_{j,l}(1 - Z_{i,j,k}))]
\]  
\[
= E[D_{i-1}](1 - E[\Pi_{j=0}^{i-1} \sum_{k=0}^{Y_i} S_{j,l}(1 - Z_{i,j,k})])
\]  
\[
= E[D_{i-1}](1 - E[\Pi_{j=0}^{i-1} \sum_{k=0}^{Y_i} S_{j,l}(1 - Z_{i,j,k})][Y, S])]
\]  
\[
= E[D_{i-1}](1 - E[\Pi_{j=0}^{i-1} \sum_{k=0}^{Y_i} S_{j,l}(1 - E[Z_{i,j,k})]])
\]  
\[
= E[D_{i-1}](1 - E[\Pi_{j=0}^{i-1} \sum_{k=0}^{Y_i} S_{j,l}(1 - E[Z_{i,j,k})]])
\]  
\[
\leq E[D_{i-1}](1 - \Pi_{j=0}^{i-1} (1 - E[Z_{i,j,k}] )^{E[\sum_{i=0}^{Y_i} S_{j,l}]})
\]  
\[
= E[D_{i-1}](1 - \Pi_{j=0}^{i-1} (1 - E[Z_{i,j,k}] )^{E[Y_j]|E[S_{i,l}]})
\]

where the third line follows as a compound distribution and the 6'th by Jensens Inequality.

\[
E[D_1] = (1 - (1 - E[Z_{i,0,1}])^C)
\]  

\[
\square
\]

**Proposition 7.** It holds:

\[
V[L_n] = E[\Omega_n] V[D_n Q_n]
\]

\[
+ E[\Omega_n](1 - E[\Omega_n]) E[D_n Q_n]^2
\]

\[
V[D_n Q_n] \approx E[D_n] V[Q_n]
\]

\[
+ E[Q_n]^2 E[D_n](1 - E[D_n])
\]

\[
V[Q_n] = E[V[Q_n|\lambda, \sigma, \mu]]
\]

\[
+ V[E[Q_n|\lambda, \sigma, \mu]]
\]

\[
V[Q_n|\lambda, \sigma, \mu] = \sum_{i=1}^{n-1} (V[Y_i|\lambda, \mu] E[S_{i,l} | \sigma]^2
\]

\[
+ E[Y_i|\lambda, \mu] V[S_{i,l} | \mu] E[Z_{i,0,1} | \mu]^2
\]

\[
+ E[Y_i|\lambda, \mu] E[S_{i,l} | \sigma] V[Z_{i,0,1} | \mu])
\]

\[
V[D_n] \approx E[D_n](1 - E[D_n])
\]

\[
V[B_n] \approx CV[Z_{n,i,k}]
\]

**Proof.** • For the variance of $L_n$ it holds:

\[
V[L_n] = E[\Omega_n]^2 V[D_n Q_n] + V[\Omega_n] E[D_n Q_n]^2 + V[\Omega_n] V[D_n Q_n]
\]

\[
= E[\Omega_n]^2 V[D_n Q_n] + E[\Omega_n](1 - E[\Omega_n]) E[D_n Q_n]^2 + E[\Omega_n](1 - E[\Omega_n]) V[D_n Q_n]
\]

\[
= E[\Omega_n] V[D_n Q_n] + E[\Omega_n](1 - E[\Omega_n]) E[D_n Q_n]^2
\]

\[
V[D_n Q_n] = E[D_n] V[Q_n] + E[Q_n]^2 E[D_n](1 - E[D_n]) + E[D_n](1 - E[D_n]) V[Q_n]
\]

\[
= E[D_n] V[Q_n] + E[Q_n]^2 E[D_n](1 - E[D_n])
\]

• For the variance of $Q_n$ it holds:

\[
V[Q_n] = E[V[Q_n|\lambda, \sigma, \mu]] + V[E[Q_n|\lambda, \sigma, \mu]]
\]

and
\[ V[Q_n | \lambda, \sigma, \mu] = \sum_{i=1}^{n-1} (V[\sum_{l=0}^{Y_i} S_{i,l}]E[Z_{n,i,1}]^2 + E[\sum_{l=0}^{Y_i} S_{i,l}]V[Z_{n,i,1}]) \]  
\[ = \sum_{i=1}^{n-1} ((V[Y_i]E[S_{i,l}]^2 + E[Y_i]E[S_{i,l}]E[Z_{n,i,1}]^2 + E[Y_i]E[S_{i,l}]V[Z_{n,i,1}])) \]  

- For \( D_n \), the inequality

\[ V[D_n] \leq E[D_n](1 - E[D_n]) \]  

follows directly from the Bhatia Davis inequality

\[ Var \leq (\text{Exp} - \min)(\max - \text{Exp}) \]  

\[ \square \]

**Proposition 8.** When \( Y_i \sim \text{Pois}(\Lambda M) \), \( \Lambda \sim \text{Gamma}(a, b) \), \( S_{i,t} \sim \text{Pois}(\Sigma) \), \( \Sigma \sim \text{Gamma}(\alpha, \beta) \), \( Z_{n,i,j} \sim \text{Bernoulli}(e^{i-n \cdot M}) \) (Bernoulli over complementary CDF of an exponential distribution), \( M \sim \text{Gamma}(a, b) \) and \( \Omega_i \sim \text{Bernoulli}(e^{i-n \cdot \Delta M}) \) it holds:

\[ E[Q_n] = \frac{\alpha \alpha + \beta \beta}{b b} ab^{\alpha}(n - i + b)^{-\alpha - 1} \]  
\[ E[D_i] \approx E[D_{i-1}](1 - \Pi_{j=0}^{i-1}(1 - (1 + \frac{i - j}{b})^{-(a)}))^{\frac{\alpha}{\alpha + \beta}} \]  
\[ E[B_n] = C(1 + \frac{n - i}{b})^{-(a)} \]  
\[ E[\Omega_n] = (1 + \frac{n}{\Delta b})^{-(a)} \]
and

\[ V[Q_n] = \sum_{i=1}^{n-1} \left[ \left( \frac{a}{b^2} + \frac{a^2}{b} \right) \left( \frac{\alpha}{\beta^2} + \frac{\alpha + \beta^2}{\beta} \right) \right] \]

\[ ab^a((2n - 2i) + b)^{-a-1} \]

\[ + \frac{a}{b} \left( \frac{\alpha}{\beta^2} + \frac{\alpha + \beta}{\beta} \right) \]

\[ (b^a(a + 1)(2n + 2i + b)) \]

\[ + \frac{a}{b} \left( \frac{\alpha}{\beta} + \frac{\alpha + \beta}{\beta} \right) \]

\[ ((n - i + b)^{-a-1} - (2n - 2i + b)^{-a-1}) \]

\[ + \frac{a^2 \alpha + \beta^2}{\beta} \sum_{1 \leq i < j < n} \]

\[ [b^a(a + 1)(2n - i - j + b)^{-a-2} \]

\[ - a^2 b^2 ((n - i + b)(n - j + b))^{-a-1}] \]

\[ + \left( \frac{\alpha^2}{\beta^2} + \frac{\alpha}{\beta^2} \right) \]

\[ \left( \sum_{i=1}^{n-1} ab^a(n - i + b)^{-a-1} \right)^2 \]

\[ + \sum_{1 \leq i < j < n} \left[ b^a(a + 1)(2n - i - j + b)^{-a-2} \right. \]

\[ - a^2 b^2 ((n - i + b)(n - j + b))^{-a-1}) \]

\[ E[S_{n,i}] = C(1 + \frac{n - i}{b})^{-a} - (1 + \frac{n - i}{b})^{-2(a)} \]

\[ V[\Omega_n] = (1 + \frac{n}{\Delta b})^{-a} - (1 + \frac{n}{\Delta b})^{-2(a)} \]

Proof: For \( Q_n \) it holds

\[ \sum_{k=1}^{n-i} Z_{n,i,k} | \lambda, \sigma, \mu = E[Y_1 | \lambda, \mu]E[S_{1,1} | \sigma]E[Z_{n,i,1} | \mu] \]

\[ = \lambda \mu (\sigma + 1)e^{(i-n)\mu} \]

\[ E[\lambda] = \frac{a}{b} \]

\[ V[\lambda] = \frac{a^2}{b^2} \]

\[ E[\sigma + 1] = \frac{\alpha + \beta}{\beta} \]

\[ V[\sigma + 1] = \frac{\alpha}{\beta^2} \]

\[ E[\mu e^{(i-n)\mu}] = \int_{0}^{\infty} \mu e^{(i-n)\mu} b^a e^{-b \mu} d\mu \]

\[ = ab^a(n - i + b)^{-a-1} \]
and thus
\[ E[Q_n] = \sum_{i=1}^{n} \frac{a}{b} \frac{\alpha + \beta}{\beta} ab^a(n - i + b)^{-a-1} \]  

(112)

Further

\[ V[Y_i|\lambda, \mu] = \lambda^2 \mu + \lambda \mu^2 + \lambda \mu \]  

(113)

\[ V[S_{1,1} \sigma] = \sigma \]  

(114)

\[ E[Z_{n,i,1}|\mu]^2 = e^{(i-n)\mu} \]  

(115)

\[ E[Z_{n,i,1}|\mu] = e^{(2i-2n)\mu} \]  

(116)

\[ = E[Z_{2n,2i,1}] \]  

(117)

\[ V[Z_{n,i,1}|\mu] = E[Z_{n,i,1}|\mu] - E[Z_{n,i,1}|\mu]^2 \]  

(118)

\[ = E[Z_{n,i,1}|\mu] - E[Z_{2n,2i,1}|\mu] \]  

(119)

\[ = ab^a((n - i + b)^{-a-1} - (2n - 2i + b)^{-a-1}) \]  

(120)

We will also need

\[ E[\lambda^2] = V[\lambda] + E[\lambda]^2 \]  

(121)

\[ = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta} \]  

(122)

\[ E[(\sigma + 1)^2] = V[\sigma + 1] + E[\sigma + 1]^2 \]  

(123)

\[ = \frac{\alpha}{\beta^2} + \frac{\alpha + \beta^2}{\beta} \]  

(124)

\[ E[\mu^2 e^{(2i-2n)\mu}] = \int_0^\infty \mu^2 e^{(2i-2n)\mu} b^a \mu^{a-1} e^{-\mu} \frac{\Gamma(1)}{\Gamma(a)} d\mu \]  

(125)

\[ = b^a \frac{\Gamma(a+2)(2n-2i+b)^{-a-2}}{\Gamma(a)} \]  

(126)

\[ = b^a(a+1)a(2n-2i+b)^{-a-2} \]  

(127)

Now we know that:

\[ E[(V[Y_i|\lambda, \mu]E[S_{1,1} \sigma]^2 E[Z_{n,i,1}|\mu]^2] = E[(\lambda^2 \mu + \lambda \mu^2 + \lambda \mu)(\sigma + 1)^2 e^{(2i-2n)\mu}] \]  

(128)

\[ = E[\lambda^2 e^{(\sigma + 1)^2} E[\mu^2 e^{(2i-2n)\mu}]] + E[\lambda e^{(\sigma + 1)^2} E[\mu^2 e^{(2i-2n)\mu}]] + E[\lambda e^{(\sigma + 1)^2} E[\mu e^{(2i-2n)\mu}]] \]  

(129)

\[ = (\frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2} + \frac{\alpha + \beta^2}{\beta}) ab^a((2n - 2i) + b)^{-a-1} \]  

(130)

\[ + \frac{\alpha}{\beta}(\frac{\alpha}{\beta^2} + \frac{\alpha + \beta^2}{\beta}) b^a(a + 1)a(2n + 2i + b)^{-a-2} \]  

(131)

\[ + \frac{\alpha}{\beta^2} + \frac{\alpha + \beta^2}{\beta}) ab^a((2n - 2i) + b)^{-a-1} \]  

(132)

\[ E[E[Y_i|\lambda, \mu]V[S_{1,1}|\mu]E[Z_{n,i,1}|\mu]^2] = E[\lambda \mu e^{(2i-2n)\mu}] \]  

(133)

\[ = E[\lambda e^{(\sigma + 1)^2} E[\mu e^{(2i-2n)\mu}]] \]  

(134)

\[ = \frac{a}{b} \frac{\alpha}{\beta^2} + \frac{\alpha + \beta^2}{\beta}) ab^a((2n - 2i) + b)^{-a-1} \]  

(135)
and

\[ E[E[Y_i|\lambda, \mu]E[S_i,t|\sigma]V[Z_{n,i,1}|\mu]] = E[\lambda \mu(\sigma + 1)E[Z_{n,i,1}|\mu] - E[Z_{2n,i,1}|\mu]] \] (138)
\[ = E[\lambda E[\sigma + 1](E[\mu e^{(i-n)\mu}] - E[\mu e^{(2i-n)\mu}])] \] (139)
\[ = \frac{a \alpha + \beta}{b} - ab^a((n - i + b)^{-a} - (2n - 2i + b)^{-a}) \] (140)

Finally for the second part of the variance we need:

\[ V[\sum_{i=1}^{n} \mu e^{(i-n)\mu}] = \sum_{1 \leq i \leq j \leq n} Cov[\mu e^{(i-n)\mu}, \mu e^{(j-n)\mu}] \] (141)
\[ = \sum_{1 \leq i \leq j \leq n} E[\mu e^{(i-n)\mu} \mu e^{(j-n)\mu}] - E[\mu e^{(i-n)\mu}]E[\mu e^{(j-n)\mu}] \] (142)
\[ = \sum_{1 \leq i \leq j \leq n} E[\mu^2 e^{(i+j-2n)\mu}] - E[\mu e^{(i-n)\mu}]E[\mu e^{(j-n)\mu}] \] (143)
\[ = \sum_{1 \leq i \leq j \leq n} \frac{b^a(\alpha + 2)(2n - i - j + b)^{-a-2}}{\Gamma(a)} \] (144)
\[ - a^2 b^2 a((n - i + b)(n - j + b))^{-a-1} \] (145)
\[ V[\lambda(\sigma + 1)] = E[\lambda^2 \sigma^2 V[\sigma + 1] + V[\lambda E[\sigma + 1]^2 + V[\lambda] V[\sigma + 1] \] (146)
\[ = \frac{a^2 \alpha + \beta^2}{b^2} + \frac{a \alpha + \beta^2}{b^2} + \frac{a \alpha}{b^2} \] (147)

With this we get

\[ V[E[Q_n|\lambda, \sigma, \mu]] = V[\sum_{i=1}^{n} E[\sum_{k=1}^{n} Z_{n,i,k}|\lambda, \sigma, \mu]] \] (148)
\[ = V[\sum_{i=1}^{n} \lambda \mu(\sigma + 1)e^{(i-n)\mu}] \] (149)
\[ = V[\lambda(\sigma + 1)] \sum_{i=1}^{n} \mu e^{(i-n)\mu}] \] (150)
\[ = E[\lambda(\sigma + 1)]^2 V[\sum_{i=1}^{n} \mu e^{(i-n)\mu}] \] (151)
\[ + V[\lambda(\sigma + 1)]E[\sum_{i=1}^{n} \mu e^{(i-n)\mu}] \] (152)
\[ + V[\lambda(\sigma + 1)] V[\sum_{i=1}^{n} \mu e^{(i-n)\mu}] \] (153)
\[ = \frac{a^2 \alpha + \beta^2}{b^2} \sum_{1 \leq i \leq j \leq n} (b^a(\alpha + 1)a(2n - i - j + b)^{-a-2} \] (154)
\[ - a^2 b^2 a((n - i + b)(n - j + b))^{-a-1} \] (155)
\[ + \frac{a^2 \alpha + \beta^2}{b^2} + \frac{a \alpha + \beta^2}{b^2} \] (156)
\[ + \frac{a \alpha}{b^2} ] \sum_{1 \leq i \leq j \leq n} (b^a(\alpha + 1)a(2n - i - j + b)^{-a-2} \] (157)
\[ - a^2 b^2 a((n - i + b)(n - j + b))^{-a-1} \] (158)

Inserting into propositions 1 and 2 now yields the result.
For $D_i$ note the following: As an exponential distribution whose rate is drawn from a Gamma distribution with shape $a$ and rate $b$ is equal to a Lomax Distribution with scale $b$ and shape $a$, a single $Z_{i,j,k}$ is equal to a Bernoulli trial over the complementary CDF of the Lomax distribution.

\[
E[Z_{i,j,k}] = (1 + \frac{i - j}{b})^{-(a)}
\]

It therefore holds

\[
E[D_i] = E[D_{i-1}] (1 - \Pi_{j=0}^{i-1} (1 - (1 + \frac{i - j}{b})^{-(a)}))^{C}
\]

\[
E[D_1] = (1 - (1 + \frac{1}{b})^{-(a)}C)
\]

$B_n$ and $\Omega_n$ follow by the same argument (Bernoulli trial over the complementary CDF of the Lomax distribution) and insertion into propositions 1 and 2.