Allocating Capacity in Parallel Queues to Improve Their Resilience to Deliberate Attack

W. Matthew Carlyle * Shane G. Henderson † Roberto Szechtman ‡

July 21, 2011

Abstract

We develop models that lend insight into how to design systems that enjoy economies of scale in their operating costs, when those systems will subsequently face disruptions from accidents, acts of nature, or an intentional attack from a well-informed attacker. The systems are modeled as parallel $M/M/1$ queues, and the key question is how to allocate service capacity among the queues to make the system resilient to worst-case disruptions. We formulate this problem as a three-level sequential game of perfect information between a defender and a hypothetical attacker. The optimal allocation of service capacity to queues depends on the type of attack one is facing. We distinguish between deterministic incremental attacks, where some, but not all, of the capacity of each attacked queue is knocked out, and zero-one random-outcome (ZORO) attacks, where the outcome is random and either all capacity at an attacked queue is knocked out or none is. There are differences in the way one should design systems in the face of incremental or ZORO attacks. For incremental attacks it is best to concentrate capacity. For ZORO attacks the optimal allocation is more complex, typically, but not always, involving spreading the service capacity out somewhat among the servers.

1 Introduction

In this paper we model decision problems where a centralized planner, also known as the *defender*, wishes to design parallel service channels (queues) with the knowledge that an *attacker* (representing worst-case disruptions, either intentional or otherwise) will attempt, in a single attack, to maximize disruption to the system. Economies of scale incentivize the defender to concentrate service capacity, which seems at odds to the goal of avoiding disruption caused by the attacker.

*Operations Research Department, Naval Postgraduate School, Monterey, CA 93943, USA. Phone: + 831-656-2106, email: mcarlyle@nps.edu

†School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853, USA. Phone: + 607-255-9126, email: sgh9@cornell.edu

‡Operations Research Department, Naval Postgraduate School, Monterey, CA 93943, USA. Phone: + 831-656-3311, email: rszechtm@nps.edu

Our model considers situations where the attacker focuses on server capacity. For example, in designing the hospital network for a community one might ask whether it is better to have two small hospitals or one large hospital. There are often economies of scale that make a single large hospital the more attractive option from the perspective of operating costs. However, when one considers the potential for disruptions such as accidents or intentional attacks that could remove some or all of a hospital's capacity, one might not be so quick to "put all the eggs in one basket." Perhaps the multiple-small-hospital solution is better in that it is more resilient to a range of disruptions. This argument might be applied recursively, leading a risk-averse planner to prefer many small hospitals to one large one. Other situations supported by our model are jamming of wireless communication networks, where the attacker increases the noise to signal ratio, de facto reducing server capacity, and physical attacks on server farms for national security data.

We model the service channels as $M/M/1$ queues, where capacity is synonymous with the service *rates* of the queues. We model the competition between the defender and attacker as a three-stage sequential game of perfect information. In the first stage, the defender chooses the initial service rates of the parallel queues. In the second stage the attacker, with full knowledge of the service rates of the parallel queues, then attacks the queues, reducing the service rates. In the third and final stage, the defender then routes traffic through the queues, attempting to maximize throughput subject to a bound on the mean steady-state volume of traffic in the system. This bound on the mean steady-state volume of traffic in the system is meant to capture the defender's desire to ensure a certain quality of service for customers of the service system, while also allowing us to capture the economies of scale in queueing systems.

Our model only considers the case in which an adversary will attack our network of queues once; if the defender believes that the attacker is capable of sustaining a campaign of repeated attacks over a long time horizon then our model is not appropriate. An analysis based on a repeated game might be slightly better suited to the situation, but we are of the opinion that, in a case where the attacker has such resources at his disposal, both the attacker and defender are probably concerned with larger, strategic issues, and that the defense of one set of servers would only constitute a small part of a much larger conflict. In such a case, the optimal allocation of scarce defensive resources among different sites would require a different (although almost certainly still game-theoretic) model.

We break up our analysis of the three-stage sequential game between defender and attacker by analyzing the last, or innermost, stage first. This is the traffic routing, or flow allocation, problem, and will be solved by the defender after the capacity allocation decision has been made by him, and also after the attacker makes any capacity reduction decision. After we determine characteristics of solutions to this innermost problem, we move one step outward to formulate the attacker's problem of determining optimal service rate reduction, and the corresponding optimal routing, for a given (fixed) capacity allocation by the defender. Finally, we analyze the entire problem of determining the optimal capacity assignment, given that for any capacity allocation an optimal capacity reducing attack, and its corresponding optimal flow allocation, will follow.

Steady-state analysis is used throughout for simplicity. Indeed, explicit expressions for the mean steady-state volume of traffic are available, and thus we can obtain a complete solution to the sequential game. If we were to attempt to use transient analysis, our results

would necessarily depend on the initial state of any queues and other very specific information that seems overly detailed given that our goal is to obtain insights from a stylized model.

Extensions to this model could include actions like repair of attacked queues. We elect not to take that path because the current model is already challenging to analyze, and reveals important insights that are the goal of the paper.

As we will show, the optimal initial allocation of capacity among the service channels by the defender depends on the *types* of disruptions, or "attacks," that are expected. We consider two types of attacks.

Incremental attacks involve a deterministic reduction in the service capacity of queues selected by the attacker. The total reduction in service capacity is constrained. Such attacks are meant to model small-scale attacks on structures that may leave the structure still partially functional. For example, a hospital may still be able to function, albeit with a limited set of services, if some but not all of its wards are rendered unusable through destruction or contamination. Similarly, a bridge may still conduct a limited traffic flow if some of its lanes are rendered unusable. For incremental attacks we find, perhaps surprisingly, that the defender's optimal first-stage decision is to concentrate all capacity in one of the queues. Upon reflection this is quite intuitive, because the defender attempts to maximize throughput subject to a bound on the mean steady-state number of customers in the system (mean occupancy). Larger-capacity queues exhibit economies of scale with respect to mean occupancy, and so all else being equal, the defender prefers to concentrate capacity. Since incremental attacks simply *reduce* capacity by a fixed amount, it is optimal for the defender to concentrate capacity.

A well-equipped attacker may be able to wreak more destruction than we model with incremental attacks, so we also consider *zero-one random-outcome* (ZORO) attacks. Here the attacker may divide attacking effort between multiple queues. With some probability, related to the attacker's effort and the capacity of the queue that is being attacked, the queue's service capacity is entirely lost, but otherwise it remains fully functional. Again the defender wishes to maximize throughput subject to a bound on the mean steady-state volume of traffic in the system. The optimal first-stage decision for the defender often involves distributing the service capacity more than for incremental attacks as intuition might suggest, although there are situations where it is again optimal to concentrate capacity.

Our work is an outgrowth of Brown et al. [2006], where 3-stage sequential games are used to analyze various potential attacks. In contrast to that work, we deal mainly with continuous decision variables and nonlinear mathematical programs. Zhuang and Bier [2007] and Golany et al. [2007] compare strategies for preparing for "probabilistic" and "strategic" threats. Probabilistic threats are those that arise from chance alone such as natural disasters, while strategic threats arise from design such as through terrorist attacks. Our analysis can be viewed as further work on strategic threats, where we distinguish different types of strategic threats, and is similar to Bier et al. [2007] in the sense that we develop insights through stylized models. The distinction between probabilistic and strategic threats has been studied for some time; see, e.g., Garrick [2002], and game theory is accepted as a useful theoretical framework to study strategic threats; see the survey Bier [2005]. The models described in Woo [2002] and Major [2002] are similar in some respects to our ZORO model, with that of Woo [2002] suggesting that low-utility targets are more attractive for attack, and that of Major [2002] suggesting that successful attacks are more likely to be on smaller targets

than larger ones. Paté-Cornell and Guikema [2002] describe a comprehensive computational framework using influence diagrams to prioritize threats and countermeasures.

Our analysis is also related to existing results in queueing theory. In particular, the third stage of our analysis of incremental attacks is similar to the flow assignment problem described in Kleinrock [1976, p. 340]. Kleinrock treats general networks, and therefore gives an algorithmic solution due to Fratta et al. [1973]. In contrast we have a (very) specific network and are therefore able to give an explicit solution. Our inner problem is also similar to Kleinrock's capacity-assignment problem where the flows are given and one wishes to choose capacities, but the capacity-assignment problem is more straightforward. The observation that some queues will remain unused in the third stage of our analysis of incremental attacks is similar to a well-known phenomenon in queues with multiple heterogeneous servers where sufficiently slow servers remain unused; see, e.g., Rubinovitch [1985] and Cabral [2005].

Relative to these earlier contributions, the key contributions of our work are that we analyze models that describe how to design parallel service systems that exhibit economies of scale, in that larger systems are more efficient than smaller systems, in the face of attacks of different types, and we show that the defender's design decisions are heavily influenced by the type of attacks they anticipate.

The remainder of this paper is organized as follows. In Section 2 we analyze deterministic incremental attacks on queues. Section 3 deals with zero-one random attacks on queues, and in Section 4 we present the conclusions of the paper.

2 Incremental Attacks on Queues

Consider n parallel $M/M/1$ queues, where the ith queue has mean arrival rate λ_i and service rate μ_i, $i = 1, \ldots, n$. Throughout, we assume that the goal of the defender, who operates the queues, is to adjust the arrival rates to the servers so as to maximize the total throughput, $\sum_i \lambda_i$, with a restriction on the total mean steady-state number of jobs in the system, Q. We can therefore define the feasible region for the arrival rate vector as $\mathcal{L} = \{\lambda_i \geq 0 : \sum_{i=1}^{n} \lambda_i/(\mu_i - \lambda_i) \leq Q\}$. If we define λ as the vector of arrival rates, the defender's decision problem is then stated as:

$$\max_{\lambda \in \mathcal{L}} \sum_{i=1}^{n} \lambda_i. \tag{1}$$

We have made the modeling choice of allowing the defender to select the service rates subject to a bound on the expected number of jobs in the system. This setup reflects the situation where the defender has considerable flexibility and control over how to direct traffic to the various queues. We impose a bound on the mean steady-state number of jobs in the system, which is similar in effect to bounding the total traffic that can be routed through the queues, but in contrast to bounding the total traffic it also reflects the economies of scale present in queueing systems.

We can rewrite the optimization problem (1) from the point of view that choosing an arrival rate for a queue is equivalent to choosing the mean steady-state number of jobs in the queue, and make a change of variables $q_i = \lambda_i/(\mu_i - \lambda_i)$ to represent the problem in terms of q, the vector of queue lengths. The feasible region for q is the simplex $\mathcal{Q} = \{q_i \geq$

$0 : \sum_{i=1}^{n} q_i \leq Q\}$, and the resulting optimization problem for the defender is:

$$\max_{q \in \mathcal{Q}} \sum_{i=1}^{n} \mu_i \frac{q_i}{1+q_i}. \tag{2}$$

We now posit an attacker who can modify the service rates at the servers, where we restrict the attacker's actions by the simple constraint

$$\sum_{i=1}^{n} \mu_i = M,$$

for some (predetermined) $M > 0$. Letting μ denote the vector of service rates, we can represent this constraint, together with the requirement that μ be nonnegative, as $\mu \in \mathcal{M}$ for an appropriately defined simplex $\mathcal{M}$. Given the attacker's selection of service rates, the defender then selects the mean number of jobs q to attempt to maximize the total throughput through the queues, subject to a bound on the mean steady-state number of jobs in the full system. The optimization problem for the attacker is therefore

$$\min_{\mu \in \mathcal{M}} g(\mu), \tag{3}$$

where

$$g(\mu) = \max_{q \in \mathcal{Q}} \quad \sum_{i=1}^{n} \mu_i \frac{q_i}{1+q_i}.$$

2.1 Solving the Inner Problem

Consider the inner maximization, where the goal is to compute $g(\mu)$ for a fixed μ. The feasible region $\mathcal{Q}$ is convex, and the objective function is strictly concave in q. Therefore, a unique global maximum exists. For now suppose that $\mu \in \mathcal{M}$ is fixed. We assume, without loss of generality, that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$.

The vector $(q^*, w^*) \in R^{n+1}$ is an optimal solution-Lagrange multiplier pair if and only if $w^* \geq 0$, $w^*(\sum_{i=1}^{n} q_i^* - Q) = 0$, and the Lagrange condition

$$q^* = \arg\max_{q_i \geq 0} \left\{ \sum_{i=1}^{n} \left(\mu_i \frac{q_i}{1+q_i} - w^* q_i \right) + w^* Q \right\}$$

holds; see pp. 490 of Bertsekas [1999] for the general theory. Due to the separable nature of the objective function, the Lagrange condition becomes

$$q_i^* = \arg\max_{q_i \geq 0} \left\{ \mu_i \frac{q_i}{1+q_i} - w^* q_i \right\}, \tag{4}$$

for $i = 1, \ldots, n$.

Solving (4) for $w \geq 0$ arbitrary results in

$$q_i(w) = \begin{cases} \left(\frac{\mu_i}{w}\right)^{1/2} - 1 & \text{if } \mu_i > w \\ 0 & \text{if } \mu_i \leq w \end{cases}, \tag{5}$$

where we define $q_i(0) = \infty$. If we assume the first k servers receive traffic, then forcing the condition $w(\sum_{i=1}^n q_i(w) - Q) = 0$ leads to

$$w(k) = \left(\frac{\sum_{i=1}^k \sqrt{\mu_i}}{Q+k} \right)^2,$$

and, defining $\nu(k) = \sum_{i=1}^k \sqrt{\mu_i}$, we get (cf. Eq. (5))

$$k^* = \max \left\{ k \geq 1 : \mu_k > \left(\frac{\nu(k)}{Q+k} \right)^2 \right\}$$

is the index of the smallest (*i.e.* slowest) server that gets any traffic. The ordering assumption of the μ_i's guarantees that k^* is well-defined.

To recap, we have $w(k^*) > 0$, $q_i(w(k^*))$ satisfies Eq. (5), and $w(k^*)(\sum_{i=1}^{k^*} q_i(w(k^*)) - Q) = 0$. We have shown the following result.

Proposition 1. *The optimal solution to the inner maximization problem is* $w^* = w(k^*)$ *and* $q_i^* = q_i(w^*)$*, where*

$$q_i^* = \begin{cases} (Q+k^*)\frac{\sqrt{\mu_i}}{\nu(k^*)} - 1 & \text{if } i \leq k^* \\ 0 & \text{if } i > k^* \end{cases}. \tag{6}$$

The optimal solution is interesting in that it does not necessarily use all of the servers. Specifically, it may ignore servers with a very low service rate. This follows because the marginal costs (measured in mean steady-state volume of traffic) are the same for all servers, and the traffic is initially allocated to servers for which the marginal gain in throughput is largest. Since the derivative with respect to q_i of the function $\mu_i q_i/(1+q_i)$ at 0 is μ_i, this means that the large servers are the first to get any traffic.

2.2 Solving the Outer Problem

If we substitute the result of Proposition 1 into the objective function of (2), we get that the optimal objective value of the inner problem is

$$g(\mu) = \sum_{i=1}^n \lambda_i(\mu) = \sum_{i=1}^{k(\mu)} \mu_i - \frac{\nu^2(k(\mu))}{Q + k(\mu)}, \tag{7}$$

where we write $\lambda(\mu)$ and $k(\mu)$ for the solution of the inner problem corresponding to a given $\mu \in \mathcal{M}$. This is a nontrivial optimization problem, because the summation index $k(\mu)$ in (7) depends on μ.

The proof is constructive, and works as follows. We start with an arbitrary solution $\mu \in \mathcal{M}$ with $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0$ and "smooth" the service rates so that the objective function continuously decreases, eventually reaching the equal-rate solution. In essence, we provide a path from any arbitrary μ to the equal-rate solution where the objective function increases along the path. The smoothing increases the minimal service rates and decreases

the maximal service rates in such a way that the sum of the rates remains constant. The smoothing continues until all of the service rates are equal.

More precisely, suppose that there are m maxima among $\mu_1, \ldots, \mu_n$, i.e., $\mu_1 = \cdots = \mu_m > \mu_{m+1}$, where $m < n$. Suppose further that there are ℓ minima among $\mu_1, \ldots, \mu_n$, i.e., $\mu_{n-\ell} > \mu_{n-\ell+1} = \cdots = \mu_n$, where $\ell > 0$. We then decrease the maximal service rates at a single rate, and increase the minimal service rates at a (possibly different) single rate so that M, the sum of the rates, remains constant. We continue this smoothing until one of two stopping conditions arises. One such condition is that the set of maxima, or minima, gets larger. In this case we simply adjust m or ℓ and continue.

We need the following lemma related to the smoothing operation, the proof of which may be found in the Appendix.

Lemma 1. *Let $x_1 \geq x_2 \geq \cdots \geq x_j \geq 0$ and let $\ell, m > 0$ be such that $\ell + m \leq j$. Let the vector y have components*

$$y_i = \begin{cases} x_i - \ell\delta & i \leq m \\ x_i & m < i \leq j - \ell \\ x_i + m\delta & j - \ell < i \leq j, \end{cases} , \tag{8}$$

so that the sum of the components in y and x are identical. Here $\delta > 0$ is small enough that the components of y remain nonnegative and ordered from largest to smallest. Let $h : [0, \infty) \to \Re$ be concave and increasing. Then

$$\sum_{i=1}^{j} h(y_i) \geq \sum_{i=1}^{j} h(x_i).$$

We now show that smoothing can only increase $k(\mu)$, the number of servers that receive non-zero traffic. The proof may be found in the Appendix.

Lemma 2. *Let $\mu \in \mathcal{M}$ with $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. Then smoothing μ in the manner described above can only increase $k(\mu)$.*

We now show that the attacker's optimal strategy is to balance the service rates. We provide a constructive proof that builds on the smoothing ideas in the Appendix. The proof shows that the optimal solution is unique and, moreover, yields the insight that the attacker can continuously improve the objective function through smoothing of the service rates.

Proposition 2. *The solution to (3) is $\mu_1 = \cdots = \mu_n = M/n$.*

Proof. Let $\mu \in \mathcal{M}$ with $\mu_1 \geq \cdots \geq \mu_n$, and suppose that the n rates are not all equal.

If $k(\mu) = n$, then smoothing leaves $k(\mu) = n$, the sum of the μ_i's remains constant, and Lemma 1 (with $j = n$) shows that $\nu(n)$ increases. Therefore, from (7) we see that g decreases until the smoothing completes with the service rates being all equal.

If $k(\mu) = m$, the number of maximal service rates, so that $\mu_1 = \mu_2 = \cdots = \mu_m$, then the objective function is given by

$$m\mu_1 - \frac{m^2\mu_1}{Q+m} = \frac{Qm\mu_1}{Q+m}$$

which decreases as μ_1 decreases, at least until $k(\mu)$ is poised to increase.

Suppose that $m < k(\mu) \leq n - \ell$, and let $k = k(\mu)$. Let $x = \mu$ and let y be the smoothed x as in (8). Then smoothing decreases the first m (maximal) components of x by $\ell\delta$, while components $m+1$ through $n-\ell$ are unchanged. Then, defining $S = \sum_{i=m+1}^{k} \sqrt{x_i}$, we see that

$$\begin{aligned} g(y) - g(x) &= -m\ell\delta + \frac{(m\sqrt{x_1} + S)^2}{Q+k} - \frac{(m\sqrt{x_1 - \ell\delta} + S)^2}{Q+k} \\ &= -m\ell\delta + \frac{m^2\ell\delta}{Q+k} + \frac{2mS\left(\sqrt{x_1} - \sqrt{x_1 - \ell\delta}\right)}{Q+k}. \end{aligned} \tag{9}$$

Now, since $x_i \leq x_1 - \ell\delta$ for $i > m$, it follows that

$$\begin{aligned} S\left(\sqrt{x_1} - \sqrt{x_1 - \ell\delta}\right) &\leq (k-m)\sqrt{x_1 - \ell\delta}\left(\sqrt{x_1} - \sqrt{x_1 - \ell\delta}\right) \\ &= (k-m)(\sqrt{x_1(x_1 - \ell\delta)} - (x_1 - \ell\delta)). \end{aligned} \tag{10}$$

The bound $\sqrt{x(x-a)} - (x-a) < a/2$ for $0 < a < x$ then shows that (10) is bounded above by $(k-m)\ell\delta/2$. This, together with (9), establishes that

$$\begin{aligned} g(y) - g(x) &\leq -m\ell\delta + \frac{m^2\ell\delta}{Q+k} + \frac{m(k-m)\ell\delta}{Q+k} \\ &= \frac{-Qm\ell\delta}{Q+k} \end{aligned}$$

establishing that g decreases while smoothing, at least until $k(\mu)$ is poised to increase.

It is possible that several arrival rates are simultaneously poised to become positive, i.e.,

$$\nu(k+j) - (Q+k+j)\sqrt{\mu_{k+j}} = 0$$

for $j = 1, \ldots, j^*$, and that this is not true for $j = j^* + 1$. From (7) and the fact that $\lambda(\mu)$ is continuous in μ we see that the function g is continuous in μ. Therefore, we can write

$$g(\mu) = \sum_{i=1}^{k+j^*} \mu_i - \frac{\nu^2(k+j^*)}{Q+k+j^*}.$$

The argument we have already given establishes that this expression decreases as the smoothing progresses. Therefore, g continues to decrease as we move through points where $k(\mu)$ increases, and this completes the proof. □

2.3 Upper Bounds on the Service Rates

In the previous sections we studied a version of our problem where the attacker could allocate service capacity arbitrarily across the n servers. In general however, the attacker attacks an established system with current service rates given by the vector $\tilde{\mu}$, say. We assume, without loss of generality, that $\tilde{\mu}_1 \geq \cdots \geq \tilde{\mu}_n$. The attack then takes the form of reducing the service rate vector. We constrain the attacker so that the sum of the service rate reductions is

some constant $\tilde{M}$. Define $M = \sum_{i=1}^{n} \tilde{\mu}_i - \tilde{M}$ and redefine the feasible region for the outer optimization to be

$$\mathcal{M} = \{0 \leq \mu \leq \tilde{\mu} : \sum_{i=1}^{n} \mu_i = M\}.$$

The difference with our previous analysis is that we now have upper bounds on the individual service rates.

The solution remains almost exactly the same as before. The attacker's optimal strategy is to attack the fastest server(s), leaving the defender with a system where the service rates are as equal as possible, modulo the upper bound constraints on μ.

Proposition 3. *The optimal solution to the upper-bounded version of the problem is to set* $\mu_i = \tilde{\mu}_i$ *for* $i > k^*$ *and*

$$\mu_1 = \mu_2 = \cdots = \mu_{k^*} = \left(M - \sum_{j=k^*+1}^{n} \tilde{\mu}_j \right) / k^*,$$

where

$$k^* = \max \left\{ 1 \leq k \leq n : \left(M - \sum_{j=k+1}^{n} \tilde{\mu}_j \right) / k \leq \tilde{\mu}_k \right\}.$$

The proof of Proposition 3 in the Appendix follows the same lines as the smoothing proof of Proposition 2, except that service rates are increased only until they hit their bounds and are then fixed.

2.4 Preparing for Incremental Attacks

Now consider the defender's initial choice of configuration $\tilde{\mu}$ of an n-server system as a *third*, outermost layer to the problem. The interpretation is as follows: The defender chooses an initial configuration $\tilde{\mu}$ with total service capacity Γ, knowing that an attacker will subsequently reduce the total service capacity to $M \leq \Gamma$ in an attempt to minimize the throughput the defender can achieve after the attack. Theorem 1 below shows that the optimal design from the defender's perspective is to concentrate capacity. The proof yields the insight that the defender can monotonically improve the three-level objective function to the optimal value from any configuration by concentrating capacity in a certain way. Let $f(\tilde{\mu})$ be the objective value associated with the solution given in Proposition 3.

Theorem 1. *The solution to the optimization problem*

$$\begin{array}{ll} \max & f(\tilde{\mu}) \\ s.t. & \sum_{i=1}^{n} \tilde{\mu}_i = \Gamma \\ & \tilde{\mu} \geq 0 \end{array}$$

is to concentrate all server capacity into one arbitrary server.

Proof. We begin from an arbitrary $\tilde{\mu}$ (assuming, without loss of generality, that $\tilde{\mu}_1 \geq \tilde{\mu}_2 \geq \cdots \geq \tilde{\mu}_n$) that is nonnegative and sums to Γ and perform a "reverse smoothing," whereby the minimal positive service rate $\tilde{\mu}_j$ is reduced to zero while $\tilde{\mu}_1$ simultaneously increases, giving $\tilde{\mu}'$ say. Proposition 3 gives the form of the optimal solutions, μ and μ' say, corresponding to $\tilde{\mu}$ and $\tilde{\mu}'$ respectively. Then $\mu'_i = 0$ for $i \geq j$. Furthermore, the maximal rates of μ that are not hard on their bounds $\tilde{\mu}$ are strictly increased in μ'. This is again a form of smoothing, and an application of our earlier results on smoothing establishes that $g(\mu) \leq g(\mu')$, i.e., $f(\tilde{\mu}) \leq \tilde{f}(\mu')$. We continue this reverse smoothing, continuously increasing f until $\tilde{\mu}_1 = \Gamma$. □

Theorem 1 establishes that the max-min-max configuration is to have a single, very large, server. This is the complete opposite of the (perhaps more intuitive) solution where the service capacity is spread equally among the n servers. To understand why this is optimal, note that the attacker attacks service capacity, attempting to equalize the service rates as much as possible. By ensuring that a single queue holds the entire service capacity, no equalization is possible, and the system capacity remains concentrated (but reduced) in a single queue. This leads to the efficiency that arises from using a single fast server as opposed to several slower servers that is a classical result in queueing theory. Essentially we are taking advantage of the economies of scale that are present in high-capacity service systems.

But what if an attacker were to completely wipe out the first server? The answer is that our formulation considered *incremental* attacks where the service capacities are reduced by a given total amount, so that completely destroying a single fast server is not possible. But in many settings, attacks *do* have the capability to destroy an entire service channel.

3 Zero-One Random-Outcome (ZORO) Attacks

Now suppose that there are n service channels with capacities given by the vector μ. The attacker again attacks these service channels, with the result that either the entire channel is destroyed (with probability p_i), or it remains untouched (with probability $1 - p_i$). The outcomes of the attacks on different service channels are independent. A fraction x_i of the attacker's total effort is allocated to service channel i, and p_i is a function $p(x_i, \mu_i)$.

The optimization problem we now consider is how the defender allocates server capacity, knowing that the attacker allocates effort in order to minimize the expected throughput and the defender gets to respond by maximizing the throughput subject to a constraint on the mean steady-state number of jobs in the system. In other words, we consider the problem

$$\begin{array}{rll} \max_{\mu \in \mathcal{M}} \min_x & E[g(\mu^r)] & \\ \text{s.t.} & \sum_{i=1}^n x_i \leq 1 & \\ & 0 \leq x_i \leq \bar{x}(\mu_i) & \forall i, \end{array} \tag{11}$$

where g is defined in (7), μ^r is the random vector of remaining capacities after the attacks, and $\bar{x}(\mu_i) = \inf\{x \geq 0 : p(x, \mu_i) = 1\}$. We can write

$$\mu^r = (\mu_1 I(U_1 > p_1), \ldots, \mu_n I(U_n > p_n)),$$

where $(U_1, \ldots, U_n)$ is a vector of i.i.d. $U(0,1)$ random variables and $I(\cdot)$ is the indicator function that equals 1 if its argument is true and 0 otherwise. Let $\mathcal{V} = \{(x_i, \mu_i) : 0 \leq x_i \leq \min\{1, \bar{x}(\mu_i)\},\ 0 \leq \mu_i \leq M\}$ be the two-dimensional set formed by all possible effort/capacity combinations. Throughout this section we make the following assumption:

(**A1**) Over $\mathcal{V}$, the function p is differentiable in x_i and μ_i, increasing as a function of x_i, and decreasing with μ_i.

Assumption (A1) is intended to capture the plausible notions that the destruction probability p should be increasing in the attacker's effort x_i and decreasing in the queue capacity μ_i. One could argue for other relationships between destruction probability and queue capacity, but this seems to be a reasonable starting point. It follows from the definition that $\bar{x}(\cdot)$ is concave in μ_i when p is a (jointly) convex function that meets assumption (A1).

Evidently, a positive resource allocation $x_i > 0$ may be optimal only if $\mu_i > 0$. Hence, when the defender has a single large server, only that server is bound to be attacked. This can be used to obtain a lower bound for Problem (11), because in this situation the defender can never have an expected throughput smaller than $(1 - p(1, M))M/(1 + 1/Q)$, where $M/(1 + 1/Q)$ is the largest allowable single-server throughput in $\mathcal{Q}$. On the other hand, the defender can never have an expected throughput larger than $M/(1 + 1/Q)$. This follows because having a single large server maximizes expected throughput when all service capacity remains intact (e.g., when the attacker only attacks servers with $\mu_i = 0$). Indeed,

$$\max_{q \in \mathcal{Q},\, \mu \in \mathcal{M}} \sum_{i=1}^{n} \mu_i/(1 + q_i^{-1}) = M \max_{q \in \mathcal{Q}} 1/(1 + q_i^{-1}),$$

for some $i = 1, \ldots, n$, and the latter equals $M/(1 + Q^{-1})$. Hence, having a single server with maximal rate M is defender-optimal when there are no attacks. We summarize these results in the next proposition.

Proposition 4. *The solution of Problem (11) is bounded by*

$$(1 - p(1, M))\frac{M}{1 + 1/Q} \leq \max_{\mu \in \mathcal{M}} \min_{x} E[g(\mu^r)] \leq \frac{M}{1 + 1/Q}.$$

When $p(1, M)$ is small, Proposition 4 shows that it is nearly optimal for the defender to concentrate its capacity into one server. However, when $p(1, M)$ is close to 1, having a single large server is not necessarily defender optimal. Consider for instance the case $p(1, M) = 1$ and $p(1/2, M/2) < 1$, as could occur when $p(x, \mu)$ is convex in both arguments. In this case

$E[g(\mu^r)] = 0$ if $\mu_1 = M$, and setting $\mu_1 = \mu_2 = M/2$ results in

$$\begin{aligned} E[g(\mu^r)] &= p(x, M/2)(1-p(1-x, M/2))\frac{M}{2}\frac{Q}{1+Q} + (1-p(x, M/2))p(1-x, M/2)\frac{M}{2}\frac{Q}{1+Q} \\ &\quad + (1-p(x, M/2))(1-p(1-x, M/2))M\frac{Q}{2+Q} \\ &\geq \frac{M}{2}\frac{Q}{1+Q}\Big[p(x, M/2)(1-p(1-x, M/2)) + (1-p(x, M/2))p(1-x, M/2) \\ &\quad + (1-p(x, M/2))(1-p(1-x, M/2))\Big] \\ &= \frac{M}{2}\frac{Q}{1+Q}\left(1-p(x, M/2)p(1-x, M/2)\right) > 0 \end{aligned}$$

for any $x \in [0, 1]$, since $p(1/2, M/2) < 1$ and (A1) imply that either $p(x, M/2) < 1$ or $p(1-x, M/2) < 1$. Therefore, the defender is better off by allocating capacity to at least two servers. We should point out, however, that in the extreme case where $p(y, yM) = 1$ for all $0 \leq y \leq 1$, the attacker destroys all servers regardless of the allocation μ chosen by the defender. This follows because setting $x_i = \mu_i/M$ leads to $p(\mu_i/M, \mu_i) = 1$ (with μ_i/M taking the place of y), and $E[g(\mu^r)] = 0$.

3.1 Simplifying ZORO: Unconstrained queue lengths

The analysis of the full ZORO case appears difficult, so to get some insight we simplify the model. Suppose that the ultimate goal is to maximize the expected *throughput* of the system, ignoring the length of the queues. In a sense, this regime arises as we let the bound on the mean steady-state number in system $q \to \infty$. It is also a reasonable approximation when all of the service rates are large, because the difference between the service rate μ_1 of a queue and the optimal flow as computed earlier is of the order $\sqrt{\mu_1}$, which is small relative to μ_1 when μ_1 is large.

Our simplified problem is then to select the service rates μ to solve the optimization problem

$$\begin{array}{rl} \max_{\mu \in \mathcal{M}} \min_x & E\left[\sum_{i=1}^n \mu_i^r\right] \\ \text{s.t.} & \sum_{i=1}^n x_i \leq 1 \\ & 0 \leq x_i \leq \bar{x}(\mu_i) \quad \forall i, \end{array} \tag{12}$$

where the inner objective function can be computed as

$$E\left[\sum_{i=1}^n \mu_i^r\right] = \sum_{i=1}^n (1-p(x_i, \mu_i))\mu_i = M - \sum_{i=1}^n \mu_i p(x_i, \mu_i),$$

which allows for a closed form solution, as we shall momentarily demonstrate.

We analyze two cases: (i) $p(x_i, \mu_i)$ is convex in x_i and μ_i over $\mathcal{V}$, and (ii) $p(x_i, \mu_i)$ is concave in x_i and μ_i over $\mathcal{V}$. In the former scenario concentrating attack resources yields higher expected damage, while diversifying attack resources is more appealing in the latter. This suggests that the defender is better off spreading out its capacity in case (i), and consolidating server capacity in case (ii).

While the scenarios about p are expressed jointly in terms of the attacker effort x_i and capacity μ_i, $p(\cdot, \mu_i)$ is a manifestation of the attacker's organizational capabilities and training, and $p(x_i, \cdot)$ depends on the physical resilience of the server (e.g., hospital or bridge) and on the kinetics of the attacking device. Therefore, when treated separately, $p(\cdot, \mu_i)$ convex in x_i appears when the attacker enjoys increasing rate of benefits (in terms of probability of destroying capacity) in his effort; this is plausible for relatively small groups of attackers, but less so for large groups. The case $p(x_i, \cdot)$ convex means that $p(x_i, \mu_i)$ grows at an increasing rate as μ_i decreases, which may take place when the server's physical resistance weakens faster than its capacity μ_i.

We start by considering the attacker's problem, using duality to obtain the optimal attacker resource allocation, as in Section 2.1. In preparation, to break symmetry and without loss of generality we impose an ordering on the service rates $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. Let $\ell = \max\{j \geq 1 : \sum_{i=1}^{j} \bar{x}(\mu_j) \leq 1\}$ be the maximum number of queues that can be knocked out completely if the attacker attacks the largest targets first, with $\ell = 0$ if $\bar{x}(\mu_1) > 1$.

The vector $(x^*, w^*) \in R^{n+1}$ is an optimal solution-Lagrange multiplier pair if and only if $w^* \geq 0$, $w^*(\sum_{i=1}^{n} x_i^* - 1) = 0$, and the Lagrange condition

$$x^* = \arg \max_{0 \leq x_i \leq \bar{x}(\mu_i)} \left\{ \sum_{i=1}^{n} (\mu_i p(x_i, \mu_i) - w^* x_i) + w^* \right\}$$

holds. Since the objective function is separable, the Lagrange condition becomes

$$x_i^* = \arg \max_{0 \leq x_i \leq \bar{x}(\mu_i)} \{\mu_i p(x_i, \mu_i) - w^* x_i\}, \quad i = 1, \ldots, n. \tag{13}$$

Proposition 5. *Suppose assumption (A1) holds. If $\sum_{i=1}^{n} \bar{x}(\mu_i) \leq 1$ then $x_i^* = \bar{x}(\mu_i)$, for $i = 1, \ldots, n$. Otherwise, if $p(x_i, \mu_i)$ is convex in x_i and μ_i over $\mathcal{V}$ and the expected damage $\mu_i\, p(x_i, \mu_i)$ is non-decreasing in μ_i, then*

$$x_i^* = \begin{cases} \bar{x}(\mu_i) & \text{for } i = 1, \ldots, \ell \\ 1 - \sum_{i=1}^{\ell} \bar{x}(\mu_i) & \text{for } i = \ell + 1 \\ 0 & \text{for } i = \ell + 2, \ldots, n \end{cases}, \tag{14}$$

where ℓ is defined as above to be the maximum number of queues that can be knocked out completely if the attacker attacks the largest targets first.

Proof. The first statement is true because the attacker gets to destroy all the servers when $\sum_{i=1}^{n} \bar{x}(\mu_i) \leq 1$, which is the best possible attack.

Otherwise, set $w^* = \mu_{\ell+1} p(x_{\ell+1}^*, \mu_{\ell+1})/x_{\ell+1}^*$ if $x_{\ell+1} > 0$, and $w^* = \mu_\ell p(x_\ell^*, \mu_\ell)/x_\ell^*$ if $x_{\ell+1} = 0$; in either case we get $0 < w^* < \infty$ and $w^*(\sum_{i=1}^{n} x_i^* - 1) = 0$. Observe that p convex means that the solution to problem (13) is an extreme point.

If $x_{\ell+1} = 0$, the assumptions about p imply that $\mu_i p(x_i, \mu_i)/x_i \leq w^*$ for any $i > \ell$ and any feasible $x_i > 0$, and $\mu_i p(x_i^*, \mu_i)/x_i^* \geq w^*$ for any $i \leq \ell$. Therefore setting $x_i = 0$ for $i > \ell$ and $x_i = \bar{x}(\mu_i)$ for $i \leq \ell$ is a solution to problem (13).

The case $x_{\ell+1} > 0$ can be handled analogously: p convex and the expected damage increasing with μ implies that $\mu_i p(x_i, \mu_i)/x_i \leq w^*$ for any $i > \ell + 1$ and any feasible $x_i > 0$, and $\mu_i p(x_i^*, \mu_i)/x_i^* \geq w^*$ for any $i \leq \ell + 1$. Hence, setting $x_i = 0$ for $i > \ell + 1$, $x_i = \bar{x}(\mu_i)$ for $i \leq \ell$, and $x_{\ell+1} = 1 - \sum_{i=1}^{\ell} \bar{x}(\mu_i)$ is a solution to problem (13). □

As a result of attacking according to (14), the defender ends up (on average) with

$$h(\mu) := E\left[\sum_{i=1}^{n} \mu_i^r\right] = \mu_{\ell+1}(1 - p(x^*_{\ell+1}, \mu_{\ell+1})) + \sum_{i=\ell+2}^{n} \mu_i \tag{15}$$

remaining server capacity. Hence, in the outermost layer of the problem the defender selects the initial configuration to maximize $h(\mu)$ over $\mu \in \mathcal{M}$ knowing that the attacker follows (14).

Theorem 2 below treats the outer layer problem. To simplify notation, let $\ell := \ell(\mu)$ and $\ell^* := \ell(\mu^*)$.

Theorem 2. *Suppose assumption (A1) holds, that $p(x_i, \mu_i)$ is convex in x_i and μ_i over $\mathcal{V}$, and that $\mu_i\, p(x_i, \mu_i)$ is non-decreasing in μ_i. Then a solution to the optimization problem (12) is $\mu_i^* = M/n$ for all i's.*

Proof. We will show that $h(\mu^*) \geq h(\mu)$ for an arbitrary server configuration $\mu \neq \mu^*$, in three different scenarios of increasing complexity: (i) $\ell = \ell(\mu) = n$; (ii) $\ell = 0$; and (iii) $0 < \ell < n$.

In Scenario (i) we have $h(\mu) = 0$ so any feasible allocation cannot further decrease h.

In Scenario (ii), Proposition 5 leads to $\bar{x}(\mu_1) > 1$ and $h(\mu) = M - \mu_1 p(1, \mu_1)$. Since $\bar{x}(\cdot)$ is concave, Jensen's inequality shows that $n\bar{x}(\mu_1^*) \geq \sum_{i=1}^{n} \bar{x}(\mu_i)$. Hence $\bar{x}(\mu_1^*) > 1/n$ and $\ell^* = \lfloor 1/\bar{x}(\mu_1^*) \rfloor < n$. We thus have

$$h(\mu^*) = M - \mu_1^* \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor - \mu_1^* p\left(1 - \bar{x}(\mu_1^*) \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor, \mu_1^*\right). \tag{16}$$

The convexity of p and the Second Fundamental Theorem of Calculus lead to

$$\begin{aligned} p(1, \mu_1) - p\left(\bar{x}(\mu_1^*) \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor, \mu_1\right) &= \int_{\bar{x}(\mu_1^*)\lfloor 1/\bar{x}(\mu_1^*)\rfloor}^{1} \frac{\partial p(x, \mu_1)}{\partial x} dx \\ &\geq \int_{0}^{1-\bar{x}(\mu_1^*)\lfloor 1/\bar{x}(\mu_1^*)\rfloor} \frac{\partial p(x, \mu_1)}{\partial x} dx \\ &= p\left(1 - \bar{x}(\mu_1^*) \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor, \mu_1\right). \end{aligned}$$

By Jensen's inequality

$$p\left(\bar{x}(\mu_1^*) \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor, \mu_1\right) \geq \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor p\left(\bar{x}(\mu_1^*), \mu_1\right)$$

Putting the last two results together

$$p(1, \mu_1) \geq \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor p\left(\bar{x}(\mu_1^*), \mu_1\right) + p\left(1 - \bar{x}(\mu_1^*) \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor, \mu_1\right). \tag{17}$$

Also,

$$\begin{aligned} \mu_1 p(\bar{x}(\mu_1^*), \mu_1) \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor + \mu_1 p\left(1 - \bar{x}(\mu_1^*) \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor, \mu_1\right) \\ \geq \mu_1^* \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor + \mu_1^* p\left(1 - \bar{x}(\mu_1^*) \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor, \mu_1^*\right) \end{aligned} \tag{18}$$

since $\mu_i\, p(x_i, \mu_i)$ is non-decreasing in μ_i and $\mu_1 \geq \mu_1^*$. Equations (17) and (18) result in $h(\mu^*) \geq h(\mu)$.

In Scenario (iii), suppose that $x^*(\mu_{\ell+1}) = 0$, so no server is attacked to less than assured destruction. Let $\tilde{\mu}_i = 1/\ell \sum_{j=1}^{\ell} \mu_j$ for $j = 1, \ldots, \ell$, and $\tilde{\mu}_i = 1/(n-\ell) \sum_{j=\ell+1}^{n} \mu_j$ for $i = \ell+1, \ldots, n$. We will show that $h(\mu^*) \geq h(\tilde{\mu}) \geq h(\mu)$. By Jensen's inequality,

$$\ell \bar{x}(\tilde{\mu}_1) \geq \sum_{i=1}^{\ell} \bar{x}(\mu_i) = 1, \tag{19}$$

the latter since $x^*(\mu_{\ell+1}) = 0$. Hence, the effort required to destroy the first ℓ servers under $\tilde{\mu}$ cannot be smaller than 1, which is the effort needed to destroy the same servers under μ, allowing us to conclude that $h(\tilde{\mu}) \geq h(\mu)$.

Another application of Jensen's inequality leads to

$$n\bar{x}(\mu_1^*) \geq \ell \sum_{i=1}^{\ell} \bar{x}(\tilde{\mu}_i) + (n-\ell) \sum_{i=\ell+1}^{n} \bar{x}(\tilde{\mu}_i).$$

We must have $(n-\ell) \sum_{i=\ell+1}^{n} \bar{x}(\tilde{\mu}_i) > 0$ because $\ell < n$ and, by Equation (19), $\bar{x}(\mu_1^*) > 1/n$. Thus $\ell^* = \lfloor 1/\bar{x}(\mu_1^*) \rfloor < n$ and (16) holds true. On the other hand Proposition 5 results in

$$h(\tilde{\mu}) = M - \tilde{\mu}_1 \left\lfloor \frac{1}{\bar{x}(\tilde{\mu}_1)} \right\rfloor - \tilde{\mu}_1 p \left(1 - \bar{x}(\tilde{\mu}_1) \left\lfloor \frac{1}{\bar{x}(\tilde{\mu}_1)} \right\rfloor, \tilde{\mu}_1 \right).$$

Jensen's inequality yields

$$\left\lfloor \frac{1}{\bar{x}(\tilde{\mu}_1)} \right\rfloor p \left(\left\lfloor \frac{\bar{x}(\tilde{\mu}_1)}{\bar{x}(\mu_1^*)} \right\rfloor \bar{x}(\mu_1^*), \tilde{\mu}_1 \right) \geq \left\lfloor \frac{1}{\bar{x}(\tilde{\mu}_1)} \right\rfloor \left\lfloor \frac{\bar{x}(\tilde{\mu}_1)}{\bar{x}(\mu_1^*)} \right\rfloor p(\bar{x}(\mu_1^*), \tilde{\mu}_1). \tag{20}$$

Having p convex and the Second Fundamental Theorem of Calculus result in

$$\begin{aligned} &\left\lfloor \frac{1}{\bar{x}(\tilde{\mu}_1)} \right\rfloor \left(1 - p \left(\left\lfloor \frac{\bar{x}(\tilde{\mu}_1)}{\bar{x}(\mu_1^*)} \right\rfloor \bar{x}(\mu_1^*), \tilde{\mu}_1 \right) \right) + p \left(1 - \bar{x}(\tilde{\mu}_1) \left\lfloor \frac{1}{\bar{x}(\tilde{\mu}_1)} \right\rfloor, \tilde{\mu}_1 \right) \\ &\quad \geq \left(\left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor - \left\lfloor \frac{1}{\bar{x}(\tilde{\mu}_1)} \right\rfloor \left\lfloor \frac{\bar{x}(\tilde{\mu}_1)}{\bar{x}(\mu_1^*)} \right\rfloor \right) p(\bar{x}(\mu_1^*), \tilde{\mu}_1) + p \left(1 - \bar{x}(\mu_1^*) \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor, \tilde{\mu}_1 \right). \end{aligned} \tag{21}$$

Adding the LHS and RHS of Equations (20–21) produces

$$\begin{aligned} &\tilde{\mu}_1 \left\lfloor \frac{1}{\bar{x}(\tilde{\mu}_1)} \right\rfloor + \tilde{\mu}_1 p \left(1 - \bar{x}(\tilde{\mu}_1) \left\lfloor \frac{1}{\bar{x}(\tilde{\mu}_1)} \right\rfloor, \tilde{\mu}_1 \right) \\ &\qquad \geq \tilde{\mu}_1 p(\bar{x}(\mu_1^*), \tilde{\mu}_1) \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor + \tilde{\mu}_1 p \left(1 - \bar{x}(\mu_1^*) \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor, \tilde{\mu}_1 \right). \end{aligned} \tag{22}$$

Like in Eq. (18),

$$\begin{aligned} &\tilde{\mu}_1 p(\bar{x}(\mu_1^*), \tilde{\mu}_1) \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor + \tilde{\mu}_1 p \left(1 - \bar{x}(\mu_1^*) \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor, \tilde{\mu}_1 \right) \\ &\qquad \geq \mu_1^* \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor + \mu_1^* p \left(1 - \bar{x}(\mu_1^*) \left\lfloor \frac{1}{\bar{x}(\mu_1^*)} \right\rfloor, \mu_1^* \right) \end{aligned} \tag{23}$$

since $\mu_i\, p(x_i, \mu_i)$ is non-decreasing in μ_i and $\tilde{\mu}_1 \geq \mu_1^*$. Putting together Equations (22–23) results in $h(\mu^*) \geq h(\tilde{\mu})$.

The proof of Scenario (iii) with $x^*(\mu_{\ell+1}) > 0$ goes along the same lines as the preceding, and is omitted. □

Thus we find that unlike the incremental attack setting discussed in Section 2, if the attacker's ability to cause damage meets the conditions of Theorem 2 then the defender is better off by diversifying server capacity. Recall that the destruction probability $p(x_i, \mu_i)$ is assumed to be decreasing in μ_i. If this probability decreases too rapidly, then the optimal solution will be to concentrate capacity. Indeed, this can happen if the destruction probability reaches zero for sufficiently small queue capacities (and therefore all larger queue capacities). Thus we require an assumption that the destruction probability does not decrease too rapidly, and so we assume that $\mu_i p(x_i, \mu_i)$ is non-decreasing in μ_i for each fixed x_i. This quantity is the expected amount of destroyed capacity in a queue with capacity μ_i for a fixed attacker effort x_i, and we are assuming this is non-decreasing in the queue capacity.

If, on the other hand, p is jointly concave in x_i and μ_i over $\mathcal{V}$, then the defender stands to gain by concentrating capacity because then the difficulty in destroying a target increases *slower* than the capacity of a target increases. This result is stated in Theorem 3 below; its proof follows along the lines of that of Theorem 2, and is consequently omitted.

Theorem 3. *Suppose assumption (A1) holds, that $p(x_i, \mu_i)$ is jointly concave in x_i and μ_i over $\mathcal{V}$, and that $\mu_i\, p(x_i, \mu_i)$ is non-increasing in μ_i. Then a solution to the optimization problem (12) is $\mu_1^* = M$ and $\mu_2^* = \cdots = \mu_n^* = 0$.*

It follows that, under the assumptions of Theorem 3, both zero-one random outcome and incremental attacks yield the same result; namely, concentrating capacity is defender-optimal.

4 Discussion

If a defender's main concern is incremental attacks, as might be the case when potential attackers are viewed as not possessing large-scale destructive capability, then concentrating capacity to take advantage of the resulting economies of scale is optimal. This sort of situation might arise in designing hospital systems for example, where small-scale attacks that could destroy some, but not all, of the service capacity of a hospital are considered far more likely than large-scale attacks that could destroy an entire hospital.

For ZORO attacks, as might arise with more capable attackers, the optimal defender design depends on whether the probability of taking down a server is convex or concave in the effort and server capacity. If it is convex (in the effort x_i and capacity μ_i) then the attacker finds it worthwhile to concentrate the attacks and the defender responds by completely distributing capacity. If, instead, the probability of succeeding to destroy a server is concave in the attacker's effort and in the server capacity, then the diminishing returns from larger attacks incentivizes the attacker to distribute their attacks, and the defender responds by concentrating capacity. So the situation with ZORO attacks is quite complex, and depends on the structure of the function $p(x_i, \mu_i)$.

One might also consider the case where p is concave and increasing in the effort x_i, and convex and decreasing in the capacity μ_i, reflecting diminishing returns to both large attacks and to concentrating capacity to discourage attacks. This case appears to be far more complex than the situations we have treated, and we have found through numerical experiments that either concentration or diversification of capacity can be optimal depending on the specific form of the functions involved.

The question of which of these forms for p is the most appropriate for modeling deliberate attacks is somewhat debatable and depends heavily on the context. For example, one might assume diminishing returns from attacker effort, so that p is concave in x. But one could also argue that an attacker needs to gather sufficient resources to mount an attack that has any chance whatsover of success, so that p is convex in the attacker's effort x. The question of which of these assumptions is most appropriate seems to depend somewhat on the capability of the attacker. For example, it seems unlikely that an attacker would possess sufficient material to make multiple radiological weapons, which would suggest that in such a setting p would be convex in attacker effort. With regard to how p depends on the service capacity μ, it seems plausible that as capacity μ increases, that p would asymptote toward zero, which would suggest that p should be convex in μ, at least eventually. However, in some settings p may not depend heavily on the size of the service capacity for small service capacities, so that p may be initially concave, and only become convex for larger capacities. Our analysis does not shed light on such situations, except when the capacities are constrained to a region of either concavity or convexity, but not both.

Acknowledgments

Matt Carlyle's research was supported by the Office of Naval Research. Shane Henderson's research was partially supported by National Science Foundation grant CMMI-0800688. Parts of this work were completed while he was visiting both the Naval Postgraduate School and the Australian National University and he thanks these institutions for their support. We would also like to thank the editorial team for their helpful comments, particularly Awi Federgruen, whose suggestions greatly improved the paper.

A Proofs

Proof of Lemma 1. For $i \leq m$, let s_i denote the (nonnegative) slope of the line segment joining $(y_i, h(y_i))$ to $(x_i, h(x_i))$, and for $i > k - \ell$ let s_i denote the (nonnegative) slope of the line segment joining $(x_i, h(x_i))$ to $(y_i, h(y_i))$. (We leave the s_i's undefined for $m + 1 \leq i \leq j - \ell$.) Since h is increasing and concave, it follows that the s_i's are increasing in i. We then

see that

$$\begin{aligned}
\sum_{i=1}^{j} h(y_i) - \sum_{i=1}^{j} h(x_i) &= \sum_{i=1}^{m} [h(y_i) - h(x_i)] + \sum_{i=j-\ell+1}^{j} [h(y_i) - h(x_i)] \\
&= -\sum_{i=1}^{m} (x_i - y_i) s_i + \sum_{i=j-\ell+1}^{j} (y_i - x_i) s_i \\
&= -\ell\delta \sum_{i=1}^{m} s_i + m\delta \sum_{i=j-\ell+1}^{j} s_i \\
&= \ell m\delta \left(\frac{1}{\ell} \sum_{i=j-\ell+1}^{j} s_i - \frac{1}{m} \sum_{i=1}^{m} s_i \right) \qquad (24) \\
&\geq 0,
\end{aligned}$$

where the inequality follows since the s_i's are increasing in i. □

Proof of Lemma 2. First note that any servers with equal service rates receive the same quantity of traffic. This follows since the function $\lambda/(\mu - \lambda)$ is convex in λ. Equally distributing any arrivals to equal-rate servers reduces the collective mean number of customers in the system. Therefore, either $k(\mu) = m$, $m < k(\mu) \leq n - \ell$, or $k(\mu) = n$, where ℓ and m are the number of servers with minimal and maximal service rates respectively.

Recall that

$$k(\mu) = \max\{1 \leq k \leq n : \nu(k) - (Q + k)\sqrt{\mu_k} < 0\}.$$

If $k(\mu) = m$, then, since $\mu_1 = \cdots = \mu_m$ throughout the smoothing operation,

$$\nu(m) - (Q + m)\sqrt{\mu_m} = -Q\sqrt{\mu_m} < 0.$$

Hence, $k(\mu)$ cannot decrease below m.

Suppose $m < k(\mu) \leq n - \ell$, and let k denote the initial value of $k(\mu)$. During the process of smoothing, and before $k(\mu)$ changes value, the rates $\mu_1, \ldots, \mu_m$ decrease, so that $\nu(k)$ decreases, while $(Q + k)\mu_k$ remains constant. It follows that $\nu(k) - (Q + k)\sqrt{\mu_k}$ remains strictly negative. Therefore, $k(\mu)$ cannot decrease below k.

Finally, suppose that $k(\mu) = n$. Let $x = \mu$ and suppose that we smooth x by δ to give y, as in (8) (with $j = n$). Let $\nu(n, x)$ be the value $\nu(n)$ evaluated at x, and define $\nu(n, y)$ similarly. Then, from (24) (with $j = n$),

$$\nu(n, y) - \nu(n, x) = \ell m\delta(s_{x_n} - s_{x_1 - \ell\delta}),$$

where s_{x_n} and $s_{x_1 - \ell\delta}$ are the slopes of chords of the function $\sqrt{\cdot}$ between x_n and $x_n + m\delta$, and $x_1 - \ell\delta$ and x_1 respectively. Now, $x_1 - \ell\delta \geq x_n + m\delta$, and so $s_{x_n} \geq s_{x_1 - \ell\delta}$. So finally,

$$\begin{aligned}
&\nu(n, y) - (Q + n)\sqrt{y_n} - [\nu(n, x) - (Q + n)\sqrt{x_n}] \\
&= \ell m\delta(s_{x_n} - s_{x_1 - \ell\delta}) - (Q + n)(\sqrt{y_n} - \sqrt{x_n}) \\
&= \ell m\delta(s_{x_n} - s_{x_1 - \ell\delta}) - (Q + n)m\delta s_{x_n} \\
&\leq m\delta(\ell - (Q + n))s_{x_n} \\
&\leq 0
\end{aligned}$$

since $\ell \leq n$ and $Q > 0$. Therefore, $\nu(n) - (Q+n)\sqrt{\mu_n}$ can only decrease during smoothing. $\square$

Proof of Proposition 3. We follow the proof of Proposition 2 closely. Recall that we assume the ordering $\tilde{\mu}_1 \geq \tilde{\mu}_2 \geq \cdots \geq \tilde{\mu}_n$ on the upper bounds. Let $\mu \in \mathcal{M}$ be an arbitrary feasible solution. We smooth μ in the usual way, uniformly decreasing the maximal values and uniformly increasing the minimial values that are not at their bounds, in such a way that the sum of the components remains fixed at M. We will show that this smoothing continuously decreases g. Eventually, we will arrive at a solution where all maximal rates are equal, and cannot be decreased further without violating the bounds on the smaller rates. Since the bounds $\tilde{\mu}_i$ are decreasing in i, it immediately follows that this is the solution given in the statement of the proposition.

Let $\mu \in \mathcal{M}$ be given. Let π denote the permutation corresponding to decreasing order of the elements of μ (breaking ties in order of $\tilde{\mu}$), so that $\mu_{\pi(1)}$ is the largest of the values in μ and $\mu_{\pi(n)}$ is the minimal value in μ. Let m be the number of maximal elements of μ, b be the number of minimal values in μ that are also at their bounds, ℓ be the number of minimal elements of μ that are not at their bounds, and $d = n - m - \ell - b$ be the number of values of μ that lie strictly between the maximal and minimal values. Hence, in decreasing order, the maximal elements are $\mu_{\pi(1)}, \ldots, \mu_{\pi(m)}$, the intermediate elements are $\mu_{\pi(m+1)}, \ldots, \mu_{\pi(m+d)}$, the minimal elements not at their bounds are $\mu_{\pi(m+d+1)}, \ldots, \mu_{\pi(m+d+\ell)}$ and $\mu_{n-b+j} = \tilde{\mu}_{n-b+j}$ for $j = 1, 2, \ldots, b$ are the minimal elements at their bounds. (The minimal elements that are at their bounds are indeed the last b elements, so that $\pi(n-b+j) = n-b+j$ for $j = 1, 2, \ldots, b$.)

Let $x = \mu$ and let y be the smoothed vector defined by

$$y_{\pi(i)} = \begin{cases} x_{\pi(i)} - \ell\delta & \text{if } 1 \leq i \leq m \\ x_{\pi(i)} & \text{if } m+1 \leq i \leq m+d \\ x_{\pi(i)} + m\delta & \text{if } m+d+1 \leq i \leq m+d+\ell, \text{and} \\ x_{\pi(i)} & \text{if } n-b+1 \leq i \leq n \end{cases}$$

for some small $\delta > 0$ that preserves the ordering π.

We consider the effects of smoothing in several cases depending on the value of $k(\mu)$, which is defined exactly as before except in terms of the ordering on μ, so that

$$k(\mu) = \max\{1 \leq k \leq n : \sum_{i=1}^{k} \sqrt{\mu_{\pi(i)}} - (Q+k)\sqrt{\mu_{\pi(k)}} < 0\}.$$

The proof of Lemma 2 goes through exactly as before for all cases where $k(\mu) \leq n - b$ so that $k(\mu)$ cannot decrease in this range. However, when $k(\mu) > n - b$, using the same steps and notation as in that proof, we find that

$$\nu(k(\mu), y) - \nu(k(\mu), x) = \ell m \delta \left(\frac{1}{\ell} \sum_{i=m+d+1}^{m+d+\ell} s_{\pi(i)} - \frac{1}{m} \sum_{i=1}^{m} s_{\pi(i)} \right) \geq 0.$$

Hence, on this range $k(\mu)$ cannot *increase*.

Consider the impact of smoothing in the different cases. If $k(\mu) = m$ then exactly as in the proof of Proposition 2, $g(\mu)$ decreases, at least until the point where $k(\mu)$ is poised to increase. The same is true for the cases $m < k(\mu) \leq m + d$ and $k(\mu) = n - b$. The remaining case is $k(\mu) > n - b$. In this case, apply Lemma 1 to the subvector $(\mu_{\pi(i)} : 1 \leq i \leq n - b)$ with function $h(\cdot) = \sqrt{\cdot}$ to conclude that $\sum_{i=1}^{n-b} \sqrt{\mu_{\pi(i)}}$ increases during the smoothing. Since the remaining μ values remain constant, it follows that $\sum_{i=1}^{k} \sqrt{\mu_{\pi(i)}}$ increases during the smoothing. From (7) (appropriately adjusted for the ordering of the μ values) we therefore conclude that $g(\cdot)$ decreases during the smoothing, at least until $k(\mu)$ is poised to decrease.

Finally, exactly as in the proof of Proposition 2, the value of g remains constant as we redefine the value of $k(\mu)$ at the point where it is poised to increase or decrease, and this completes the proof. □

References

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.

V. Bier. Game-theoretic and reliability methods in counter-terrorism and security. In A. Wilson, N. Limnios, S. Keller-McNulty, and Y. Armijo, editors, *Modern Statistical and Mathematical Methods in Reliability*, volume 10 of *Series on Quality, Reliability and Engineering Statistics*, pages 17–28. World Scientific, 2005.

V. Bier, S. Oliveros, and L. Samuelson. Choosing what to protect: Strategic defensive allocation against an unknown attacker. *Journal of Public Economic Theory*, 9:563–587, 2007.

G. Brown, M. Carlyle, J. Salmerón, and K. Wood. Defending critical infrastructure. *Interfaces*, 36(6):530–544, 2006.

F. B. Cabral. The slow server problem for uninformed customers. *Queueing Systems*, 50: 353–370, 2005.

L. Fratta, M. Gerla, and L. Kleinrock. The flow deviation method — an approach to store-and-forward communication network design. *Networks*, 3:97–133, 1973.

B. J. Garrick. Perspectives on the use of risk assessment to address terrorism. *Risk Analysis*, 22(3):421–423, 2002.

B. Golany, E. H. Kaplan, A. Marmur, and U. G. Rothblum. Nature plays with dice – terrorists do not: Allocating resources to counter strategic versus probabilistic risks. *European Journal of Operational Research*, 2007. doi: 10.1016/j.ejor.2007.09.001. To appear.

L. Kleinrock. *Queueing Systems Volume 2: Computer Applications*. Wiley, New York, 1976.

J. A. Major. Advanced techniques for modeling terrorism risk. *Journal of Risk Finance*, 4: 15–24, 2002.

E. Paté-Cornell and S. Guikema. Probabilistic modeling of terrorist threats: A systems analysis approach to setting priorities among countermeasures. *Military Operations Research*, 7(4):5–23, 2002.

M. Rubinovitch. The slow server problem. *Journal of Applied Probability*, 22(1):205–213, 1985.

G. Woo. Quantitative terrorism risk assessment. *Journal of Risk Finance*, 4:7–14, 2002.

J. Zhuang and V. Bier. Balancing terrorism and natural disasters — defensive strategy with endogenous attacker effort. *Operations Research*, 55:976–991, 2007.