File Fragmentation over an Unreliable Channel

Jayakrishnan Nair[1], Martin Andreasson[2], Lachlan L. H. Andrew[3], Steven H. Low[1], and John C. Doyle[1]

[1]Engineering and Applied Science, California Institute of Technology, USA
[2]Optimization and Systems Theory, Royal Institute of Technology, Sweden
[3]Centre for Advanced Internet Architectures, Swinburne University of Technology, Australia

***Abstract*—It has been recently discovered that heavy-tailed file completion time can result from protocol interaction even when file sizes are light-tailed. A key to this phenomenon is the RESTART feature where if a file transfer is interrupted before it is completed, the transfer needs to restart from the beginning. In this paper, we show that independent or bounded fragmentation guarantees light-tailed file completion time as long as the file size is light-tailed, i.e., in this case, heavy-tailed file completion time can only originate from heavy-tailed file sizes. If the file size is heavy-tailed, then the file completion time is necessarily heavy-tailed. For this case, we show that when the file size distribution is regularly varying, then under independent or bounded fragmentation, the completion time tail distribution function is asymptotically upper bounded by that of the original file size stretched by a constant factor. We then prove that if the failure distribution has non-decreasing failure rate, the expected completion time is minimized by dividing the file into equal sized fragments; this optimal fragment size is unique but depends on the file size. We also present a simple blind fragmentation policy where the fragment sizes are constant and independent of the file size and prove that it is asymptotically optimal. Finally, we bound the error in expected completion time due to error in modeling of the failure process.**

I. Motivation and summary

It has been recently discovered that heavy-tailed job completion time can result from protocol interaction even when the job size is light-tailed, provided its distribution has infinite support [1]–[4]. Indeed, the completion time can be heavy-tailed even when the job size has a tail that decays exponentially or superexponentially. A key to this phenomenon is the RESTART feature where if a job is interrupted in the middle of its processing, the entire job needs to restart from the beginning, i.e., the work that is partially completed is lost. This can model, e.g., a packet that is corrupted by bit errors needs to be retransmitted. This effect has been shown to be robust to several schemes aimed at alleviating it. The fragmentation scheme of [5], which uses the sizes of the previous $k + m$ server availability periods, lightens the completion time tail by adding k additional moments, but the resulting tail is still heavy. Multipath is explored in [6] to mitigate power-law completion time. It is shown there that redundant routing, where the entire file is sent along multiple paths and the completion time is the time when the first copy arrives at the destination correctly, preserves the power law. Split routing, where disjoint fragments of the file are sent along multiple paths and the completion time is the time when the last fragment arrives, also retains a power-law completion time though the tail can be lightened with a larger index.

In this paper, we show that the heavy-tailed completion times can actually be quite fragile and are removed by a large class of fragmentation schemes. In particular, we consider a model for file transfer over an unreliable channel and propose fragmentation policies that guarantee light-tailed completion time for light-tailed file sizes. In the models of [1]–[4], heavy-tailed completion time seems to arise from repeated comparison of a sequence of independent, identically distributed (i.i.d.) random variables (availability periods) with the *same* random variable (original job size) that has an *infinite support*. This motivates fragmentation policies that avoid this character. Our goal is to exhibit two classes of such policies, analyze the tail distribution of the completion times under these policies, and study the optimal policy that minimizes the expected completion time.

Specifically, we consider policies that partition files into fragments with independent, or bounded sizes; note that packet sizes are naturally bounded by network hardware. We show that these policies produce a light-tailed completion time as long as the original file size is light-tailed, i.e., in this case, a heavy-tailed file completion time can only originate from a heavy-tailed file size (Section III). If the file size is heavy-tailed, then the file completion time is necessarily heavy-tailed. In this case, we show that if the file size distribution is regularly varying, then under independent or bounded fragmentation, the completion time tail distribution function is asymptotically upper bounded by that of the original file size stretched by a constant factor. This means that in the degree sense, the completion time distribution is only as heavy-tailed as the job size distribution. This naturally raises the question of optimal fragmentation that minimizes the expected job completion time. We prove that if the failure distribution has non-decreasing failure rate, it is optimal to divide the file into equal sized fragments, whose size depends on the file size (Section IV-A). We also present a simple blind fragmentation policy where the fragment size is constant and independent of the file size and prove that its expected file completion time is asymptotically optimal (Section IV-B). The optimal policy as well as the suboptimal blind policy create bounded fragments, and therefore produce desirable completion time tail behavior, as described above (Section IV-C). Finally, we present simple bounds on the error in expected completion time when there is error in modeling the failure process (Section V).

II. Model and preliminaries

A. Model

Consider a file with a possibly random size $L > 0$. The file is fragmented into packets which are then sent over an unreliable channel with unit transmission rate. A packet contains a fragment of the file and a fixed-sized overhead (header, trailer). The larger the packet size, the more likely the transmission is to fail. This will be the case, e.g., if the channel randomly introduces independent bit errors so a packet with more bits has a higher probability of being corrupted and needing a retransmission; see [7, p. 132] for such a failure model for satellite and terrestrial communications. More generally, for the nth transmission attempt, let $x_n + \phi$ be the packet size, where x_n is the size of the file fragment and ϕ is the constant overhead. All sizes are measured in terms of the transmission time over the channel with unit rate. Let $(A_n, n = 1, 2, \dots)$ be i.i.d. non-negative random variables with common distribution F and independent of L, with $P(A_1 > \phi) > 0$. The nth transmission attempt will be successful if and only if $A_n \geq x_n + \phi$.

To formulate the problem precisely, we abuse notation and use $x = (x_n, n = 1, 2, \dots)$ to denote both the control (fragmentation) policy and the fragment sizes under the policy, depending on the context. Let the state $l_n := l_n^x$ be the remaining file size just after the start of the nth transmission under control policy x. Then the state l_n evolves according to,

$$l_{n+1} = l_n - x_n \, \mathbf{1}(A_n \geq x_n + \phi), \qquad n = 1, 2, \dots \quad (1)$$

$$l_1 = L \quad (2)$$

where $\mathbf{1}(z) = 1$ if z is true and 0 otherwise. We implicitly restrict ourselves to admissible policies x under which $0 \leq x_n \leq l_n$ for all n. We emphasize that the state sequence $(l_n, n \geq 1)$ depends on the control policy $x = (x_n, n \geq 1)$ though this is not explicit in the notation. The time between the nth and the $n+1$st submission is the cost at the nth stage and is given by:

$$\tau_n := (x_n + \phi) \, \mathbf{1}(l_n > 0) \quad (3)$$

Clearly, the transmission time sequence $(\tau_n, n \geq 1)$ also depends on the control x. Let $T(L)$ be the file completion time under control x as a function of the initial file size L;

$$T(L) := T^x(L) := \sum_{n \geq 1} \tau_n. \quad (4)$$

In summary, our file fragmentation model is specified by (1)–(4) with the i.i.d. random sequence $(A_n, n \geq 1)$. In subsequent sections, we will study the impact of the choice of the fragment sizes $(x_n, n = 1, 2, \dots)$ on the file completion time.

Our model is an adaptation of the model in [1]–[4] where a server alternates between availability periods and unavailability periods. There, the server availability periods have durations $(A_n, n \geq 1)$ that are i.i.d. random variables. The unavailability periods have durations $(U_n, n \geq 1)$ that are i.i.d. and independent of $(A_n, n \geq 1)$. Without fragmentation, the entire file is submitted at the beginning of each availability period until it completes successfully, $x_n = L$ for all n. Our model here has $U_n = 0$; furthermore, the one-stage cost is $x_n + \phi$ in our case but A_n (before successful transmission) in theirs. This models the case where the sender is informed of the failure only after the entire packet has been sent. These differences do not qualitatively change our conclusions (see a parallel set of results in [8] for a *job* fragmentation model that is closer to the model in [1]–[4] and the models in the checkpointing literature reviewed in Section VI).

B. Notation and preliminaries

Throughout this paper, $\overline{\lim}$ denotes the limit superior, $\underline{\lim}$ the limit inferior and $\mathbb{E}[\cdot]$ the expectation. For any functions $\gamma(t)$ and $\lambda(t)$,

1) $\gamma(t) \sim \lambda(t)$ means $\lim_{t\to\infty} \gamma(t)/\lambda(t) = 1$,
2) $\gamma(t) \lesssim \lambda(t)$ means $\overline{\lim}_{t\to\infty} \gamma(t)/\lambda(t) \leq 1$,
3) $\gamma(t) = o(\lambda(t))$ means $\lim_{t\to\infty} \gamma(t)/\lambda(t) = 0$.

Consider non-negative random variables X, Y. We will use the notation $X \leq_{\text{a.s.}} Y$ to mean $X \leq Y$ almost surely. The notation $X \leq_{\text{st}} Y$ means X is stochastically dominated by Y, i.e., $P(X > t) \leq P(Y > t)$ for all $t \geq 0$. It is easy to see that $X \leq_{\text{a.s.}} Y$ implies $X \leq_{\text{st}} Y$.

Lemma 1. *If random variables A, B, C satisfy $A \leq_{\text{st}} B \leq_{\text{st}} C$, and $P(A > x) \sim P(C > x)$, then*

$$P(A > x) \sim P(B > x) \sim P(C > x).$$

The elementary proof is omitted. Let $G(x) = P(X \leq x)$ denote the distribution function (df) of non-negative random variable X and $\overline{G}(x) := 1 - G(x)$ denote its tail df.

Definition 1. *The df G (or the random variable X) is said to be* heavy-tailed (HT) *if $\overline{\lim}_{x\to\infty} e^{\theta x}\overline{G}(x) = \infty$ for all $\theta > 0$. The df G (or the random variable X) is said to be* light-tailed (LT) *if it is not HT, i.e., if there exists a $\theta > 0$ such that $\lim_{x\to\infty} e^{\theta x}\overline{G}(x) = 0$.*

Intuitively, a distribution is HT if its tail df is (asymptotically) heavier than that of any exponential distribution. Conversely, a distribution is LT if its tail df is (asymptotically) dominated by that of some exponential distribution. The following lemma describes some closure properties of the class of LT distributions we will use in this paper.

Lemma 2. *[Closure properties of LT distributions]*

1) *Let X, Y be non-negative random variables satisfying $X \leq_{\text{st}} Y$. If Y is LT, then X is LT.*
2) *Let X, Y be non-negative random variables. If X, Y are LT, then $X + Y$ is LT.*
3) *Let $(X_i, i \geq 1)$ be a sequence of non-negative i.i.d. LT random variables, and N be an integer random variable. If N is LT, then the random sum $\sum_{i=1}^{N} X_i$ is LT.*
4) *Let L be a non-negative random variable and $\{X_i\}_{i\geq 1}$ a sequence of non-negative i.i.d. random variables independent of L and satisfying $P(X_i > 0) > 0$. If L is LT, so is $\inf\{n | \sum_{i=1}^{n} X_i \geq L\}$.*

We omit the proof of this lemma due to space limitation. An important class of HT distributions is the class of regularly varying distributions (see [9], Chapter 2 of [10]).

Definition 2. *A df G is regularly varying with index/degree $\alpha > 0$ (denoted $G \in \mathcal{RV}(\alpha)$) if*

$$\overline{G}(x) = x^{-\alpha}\chi(x)$$

where $\chi(x)$ is a slowly varying function, i.e., $\chi(x)$ satisfies

$$\lim_{x\to\infty} \frac{\chi(xy)}{\chi(x)} = 1 \quad \forall y > 0.$$

We will abuse notation and use $L \in \mathcal{RV}(\alpha)$ to mean the df G_L of a random variable L is in $\mathcal{RV}(\alpha)$. Regularly varying distributions are a generalization of the class of Pareto distributions, also referred to as power-law distributions or Zipf distributions. The closer α is to 0, the 'heavier' the tail df is.

Lemma 3. *Consider non-negative random variables X, Y. If $X \in \mathcal{RV}(\alpha)$ and $P(X > t) \sim P(Y > t)$, then $Y \in \mathcal{RV}(\alpha)$.*

The proof follows easily from the definition.

Lemma 4. *If $X \in \mathcal{RV}(\alpha)$, then $P(X > t) \sim P(X > t + c)$ for all $c \in \mathbb{R}$.*

This lemma is a consequence of the fact that regularly varying distributions are a sub-class of the class of long-tailed distributions; see [11].

Lemma 5. *If $\chi(x)$ is slowly varying, then*

$$\lim_{x\to\infty} x^{\beta}\chi(x) = \begin{cases} \infty & \text{if } \beta > 0 \\ 0 & \text{if } \beta < 0 \end{cases}.$$

See Prop. 2.6 in [10] for a proof.

III. Completion time tail asymptotics

In this section, we study the tail behavior of the completion time under a broad class of fragmentation policies. To motivate our results, we first state the following theorem, which considers the case of no fragmentation.

Theorem 6 ([1]–[4])**.** *Without fragmentation, i.e., $x_n = L$ for all n, $T(L)$ is HT as long as L has infinite support.*

The proof follows from Lemma 1 in [4]. Theorem 6 implies that without fragmentation, the completion time $T(L)$ can be HT even for LT file sizes, e.g., file size distributions with an exponential or even superexponential tail df. Our results in this section (Theorems 7–9) imply that under a broad class of fragmentation policies, the completion time $T(L)$ is LT provided L is LT. Thus, with these policies, *heavy-tailed completion times can only arise from heavy-tailed file sizes.* Moreover, we show if L is HT (specifically, regularly varying), then the tail df of $T(L)$ is bounded above by a scaled version of the tail df of L. This means that in the degree sense, the completion time is only as heavy-tailed as the file size.

A. Results

We now define the three classes of fragmentation policies studied in this section.

- **Independent fragmentation:** $x_n = \min\{X_n, l_n\}, n \geq 1$, where $(X_n, n \geq 1)$ is a sequence of i.i.d. strictly positive light-tailed random variables independent of L and $(A_n, n \geq 1)$ such that $P(A_1 \geq X_1 + \phi) > 0$.
- **Bounded fragmentation:** x_n satisfies $\min\{b, l_n\} \leq x_n \leq \min\{c, l_n\}$, $n \geq 1$, for some constants $0 < b \leq c$ such that $P(A_1 \geq c + \phi) > 0$.
- **Constant fragmentation:** $x_n = \min\{b, l_n\}$ for some constant $b > 0$ satisfying $P(A_1 \geq b + \phi) > 0$. This is a special case of independent fragmentation and of bounded fragmentation.

We now state our results for each of these classes.

Theorem 7 (Independent fragmentation)**.** *Under the independent fragmentation policy*

1) *If L is light-tailed, then $T(L)$ is light-tailed.*
2) *If $L \in \mathcal{RV}(\alpha)$, then $P(T(L) > t) \lesssim P\left(L > \frac{t}{\sigma}\right)$ where*

$$\sigma = \frac{\mathbb{E}[X_1] + \phi}{P(X_1 + \phi \leq A_1)\,\mathbb{E}[X_1 | X_1 + \phi \leq A_1]}.$$

The next result says that any policy that does not choose arbitrarily large or arbitrarily small fragment sizes produces LT completion time provided L is LT.

Theorem 8 (Bounded fragmentation)**.** *Under the bounded fragmentation policy*

1) *If L is light-tailed, then $T(L)$ is light-tailed.*
2) *If $L \in \mathcal{RV}(\alpha)$, then $P(T(L) > t) \lesssim P\left(L > \frac{t}{\sigma}\right)$ where*

$$\sigma = \frac{c + \phi}{bP(A_1 \geq c + \phi)}.$$

Intuitively, if packet size is too small, the overhead can dominate the transmission, reducing efficiency. If the packet is too large, the failure probability can be too high. Hence it is reasonable to choose packet sizes that are neither too small nor too large. Theorem 8 then guarantees that any reasonable fragmentation policy 'lightens' the completion time tail.

Since constant fragmentation is a special case of independent and bounded fragmentation, Theorems 7 and 8 imply that under constant fragmentation, $T(L)$ is LT if L is LT. When L is regularly varying, we have a sharper characterization of the asymptotics: $T(L)$ is regularly varying with the same degree.

Theorem 9 (Constant fragmentation)**.** *Under the constant fragmentation policy*

1) *If L is light-tailed, then $T(L)$ is light-tailed.*
2) *If $L \in \mathcal{RV}(\alpha)$, then $P(T(L) > t) \sim P\left(L > \frac{t}{g(b)}\right)$ where*

$$g(x) = \frac{x + \phi}{xP(A_1 \geq x + \phi)}.$$

Theorem 9 motivates choosing the constant fragment size $a := \arg\min_{x>0} g(x)$. Within the class of constant fragmentation policies, this choice produces in some sense the lightest

possible completion time tail asymptotics. We will prove in Section IV that this policy also almost minimizes the expected completion time; see Theorem 15.

B. Proofs of Theorems 7–9

Proofs of Theorems 7–9 rely on Lemma 10, which we state and prove first.

Lemma 10. *Let L be a random variable, and $(X_n, n \geq 1)$ be a sequence of i.i.d. strictly positive LT random variables independent of L and $(A_n, n \geq 1)$ such that $P(A_1 > X_1 + \phi) > 0$. Let*

$$\begin{aligned} Y_n &:= X_n \mathbf{1}(X_n + \phi \leq A_n), \\ M &:= \inf\left\{m : \sum_{n=1}^{m} Y_n \geq L\right\}, \qquad (5) \\ \tilde{T}(L) &:= \sum_{n=1}^{M} (X_n + \phi). \qquad (6) \end{aligned}$$

1) *If L is LT, then $\tilde{T}(L)$ is LT.*
2) *If $L \in \mathcal{RV}(\alpha)$, then $P\left(\tilde{T}(L) > t\right) \sim P(L > t/\sigma)$ where*

$$\sigma = \frac{\mathbb{E}[X_1] + \phi}{P(X_1 + \phi \leq A_1)\,\mathbb{E}[X_1 | X_1 + \phi \leq A_1]}.$$

The proof of this lemma for the case of regularly varying L is based on the following theorem, proved in [12].

Theorem 11 ([12])**.** *Let $L \in \mathcal{RV}(\alpha)$. For $t \geq 0$, let $R(t)$ be a non-negative, almost surely non-decreasing stochastic process independent of L satisfying the following conditions:*

1) *For some $\gamma \in (0,1)$, $\lim_{t\to\infty} R(t)/t = \gamma$ a.s..*
2) *For some positive finite constant K, $P(R(t)/t < K) = o(P(L > t))$.*

Then $P(L > R(t)) \sim P(L > \gamma t)$.

Proof of Lemma 10: We consider the cases of LT and regularly varying L separately.

Case 1: L is LT. Under the assumptions of the lemma, $(Y_n, n \geq 1)$ is an i.i.d. sequence satisfying $P(Y_1 > 0) > 0$. Invoking Lemma 2(4), we conclude from (5) that M is LT. It follows that $\tilde{T}(L)$ is LT from (6) invoking Lemma 2(3).

Case 2: $L \in \mathcal{RV}(\alpha)$. Let $N(t) := \sup\{n : \sum_{i=1}^{n}(X_i + \phi) \leq t\}$, $R(t) := \sum_{i=1}^{N(t)} Y_i$. Note that $P\left(\tilde{T}(L) > t\right) = P(R(t) < L)$. To complete the proof, it suffices to show that the process $R(t)$ satisfies conditions (1) and (2) of Theorem 11 with $\gamma = 1/\sigma$.

Condition (1) of Theorem 11 is verified using the renewal reward theorem.

$$\lim_{t\to\infty} \frac{R(t)}{t} = \frac{\mathbb{E}[Y_1]}{\mathbb{E}[X_1 + \phi]} = \frac{1}{\sigma}$$

almost surely. Note that $\sigma > 1$ since $\phi > 0$. To verify Condition (2), pick $K \in (0, 1/\sigma)$. Since $K < 1/\sigma$, we can find $\eta, \nu > 0$ such that $K = \eta\nu$, $\eta < \mathbb{E}[Y_1]$ and $\nu < 1/\mathbb{E}[X_1 + \phi]$. Then

$$\begin{aligned} P(R(t) < Kt) &= P\left(\sum_{i=1}^{N(t)} Y_i < Kt\right) \\ &= P(N(t) < t\nu) - P\left(\sum_{i=1}^{N(t)} Y_i \geq Kt \wedge N(t) < t\nu\right) \\ &\quad + P\left(\sum_{i=1}^{N(t)} Y_i < Kt \wedge N(t) \geq t\nu\right) \\ &\leq P(N(t) < t\nu) + P\left(\sum_{i=1}^{N(t)} Y_i < Kt \wedge N(t) \geq t\nu\right) \\ &\leq P\left(\sum_{i=1}^{\lfloor t\nu \rfloor} (X_i + \phi) \geq t\right) + P\left(\sum_{i=1}^{\lceil t\nu \rceil} Y_i < Kt\right) \\ &\leq P\left(\sum_{i=1}^{\lfloor t\nu \rfloor} (X_i + \phi) \geq \frac{\lfloor t\nu \rfloor}{\nu}\right) + P\left(\sum_{i=1}^{\lceil t\nu \rceil} Y_i < \eta \lceil t\nu \rceil\right). \end{aligned}$$

Noting that $1/\nu > \mathbb{E}[X_1 + \phi]$ and $\eta < \mathbb{E}[Y_1]$, and that X_1, Y_1 are LT, we can use the Chernoff bound to argue that there exist positive constants C, λ such that for large enough t,

$$P(R(t) < Kt) \leq Ce^{-\lambda t}.$$

Since $P(L > t) = t^{-\alpha}\chi(t)$ for slowly varying χ, this implies

$$\overline{\lim_{t\to\infty}} \frac{P(R(t) < Kt)}{P(L > t)} \leq \overline{\lim_{t\to\infty}} \frac{Ce^{-\lambda t}}{t^{-\alpha}\chi(t)} = \overline{\lim_{t\to\infty}} \frac{Ct^{\alpha+1}e^{-\lambda t}}{t\chi(t)} = 0.$$

The last step above uses Lemma 5. It follows that $P(R(t) < Kt) = o(P(L > t))$. This completes the proof. ■

We are now ready to prove Theorems 7–9.

Proof of Theorem 7: Consider the completion time $\tilde{T}(L)$ under the policy $\tilde{x}_n := X_n$. Clearly $T(L) \leq_{\text{a.s.}} \tilde{T}(L)$.

If L is LT, then from Lemma 10, we conclude that $\tilde{T}(L)$ is LT, which implies $T(L)$ is LT (Lemma 2(1)).

If $L \in \mathcal{RV}(\alpha)$, then from Lemma 10, we conclude that $P\left(\tilde{T}(L) > t\right) \sim P\left(L > \frac{t}{\sigma}\right)$. Since $T(L) \leq_{\text{a.s.}} \tilde{T}(L)$, it follows that $P(T(L) > t) \lesssim P\left(L > \frac{t}{\sigma}\right)$. ■

Proof of Theorem 8: Define $\tilde{L} := cL/b$. With file size $\tilde{L}$, consider the policy $\tilde{x}_n = \min\{c, \tilde{l}_n\}$, $n \geq 1$, where $\tilde{l}_1 = \tilde{L}$, $\tilde{l}_n$ denotes the remaining file size just after the nth submission. Note that this policy satisfies the conditions of Theorem 7. Denote the completion time under this scheme by $T^c(\tilde{L})$.

We will now argue that $T(L) \leq_{\text{a.s.}} T^c(\tilde{L})$. Consider a sample path, determined by the realization of L, $(A_n, n \geq 1)$ and the fragment sizes $(x_n, n \geq 1)$. Noting that for any n, if fragment submission $\tilde{x}_n$ succeeds, then submission x_n succeeds, it can be seen that $l_n \leq b\tilde{l}_n/c$ for all $n \geq 1$. This implies $T(L) \leq T^c(\tilde{L})$.

If L is LT, so is $\tilde{L}$. Theorem 7 then implies that $T^c(\tilde{L})$ is LT, which implies $T(L)$ is LT (Lemma 2(1)).

If $L \in \mathcal{RV}(\alpha)$, it is easy to see that $\tilde{L} \in \mathcal{RV}(\alpha)$. Theorem 7 implies that

$$\begin{aligned} P\left(T^c(\tilde{L}) > t\right) &\lesssim P\left(\tilde{L} > \frac{tcP\left(A_1 \geq c+\phi\right)}{c+\phi}\right) \\ &= P\left(L > \frac{tbP\left(A_1 \geq c+\phi\right)}{c+\phi}\right) \\ &= P\left(L > \frac{t}{\sigma}\right). \end{aligned}$$

Since $T(L) \leq_{\text{a.s.}} T^c(\tilde{L})$, we conclude that $P\left(T(L) > t\right) \lesssim P\left(L > \frac{t}{\sigma}\right)$. ∎

Proof of Theorem 9:

Since constant fragmentation is a special case of independent and bounded fragmentation, the proof for the case of LT L follows directly from Theorems 7 or 8.

Assume then that $L \in \mathcal{RV}(\alpha)$. We will invoke Lemma 10 with $X_n := b$, $n \geq 1$. Define

$$\hat{L} := b\left\lfloor \frac{L}{b} \right\rfloor, \quad \tilde{L} := b\left\lceil \frac{L}{b} \right\rceil.$$

It is easy to see that

$$\tilde{T}(\hat{L}) \leq_{\text{a.s.}} T(L) \leq_{\text{a.s.}} \tilde{T}(\tilde{L}).$$

We will now argue that $\hat{L}, \tilde{L} \in \mathcal{RV}(\alpha)$. Clearly,

$$\max\{L-b, 0\} \leq_{\text{a.s.}} \hat{L} \leq_{\text{a.s.}} L \leq_{\text{a.s.}} \tilde{L} \leq_{\text{a.s.}} L+b.$$

Using Lemma 4, we see that $P\left(\max\{L-b,0\} > t\right) \sim P\left(L+b > t\right)$. This implies, using Lemma 1, that

$$P\left(\hat{L} > t\right) \sim P\left(L > t\right) \sim P\left(\tilde{L} > t\right),$$

which in turn implies $\hat{L}, \tilde{L} \in \mathcal{RV}(\alpha)$ (see Lemma 3). By Lemma 10, we see that

$$P\left(\tilde{T}(\hat{L}) > t\right) \sim P\left(\tilde{T}(\tilde{L}) > t\right) \sim P\left(L > \frac{t}{g(b)}\right).$$

This implies $P\left(T(L) > t\right) \sim P\left(L > \frac{t}{g(b)}\right)$ by Lemma 1. ∎

IV. Optimal fragmentation

In the previous section, we studied the tail asymptotics of the completion time; in this section, we turn our attention to its mean. Specifically, under the assumption that F has a non-decreasing failure rate, we derive the fragmentation policy that minimizes the expected completion time. We show that this policy divides the file into equal sized fragments, whose size depends on the file size. We also present a fragmentation policy that is blind to the file size, but is asymptotically optimal. We show that under both these policies, the completion time is LT so long as L is LT. If L is regularly varying, then the completion time is regularly varying with the same index.

Consider

$$\min_x \ \mathbb{E}\left[T^x(L)\right] := \min_x \ \left(\lim_{N\to\infty} \ \mathbb{E}\left[\sum_{n=1}^{N} \tau_n \mid l_1 = L\right]\right) \quad (7)$$

An *optimal policy* is one that achieves the minimum of (7). We will restrict ourselves to the class of stationary Markov policies where the decision at time n depends only on the state l_n and not on the time n nor on past states. Since any optimal policy will never choose fragment sizes x_n with $P(A_1 \geq x_n + \phi) = 0$, we will assume without loss of generality that $P(A_1 \geq x_n + \phi) > 0$ for the class of policies that we consider. Our discussion in this section (except in IV-C, which deals with completion time tail asymptotics) will be for some generic realization of the initial file size $L > 0$.

A. Optimal policy

A stationary Markov policy is a function $x(l)$ of the remaining file size l with the following interpretation. Given l, a packet of size $x(l) + \phi$ is formed. If the packet is successfully transmitted, the remaining file size will be $l - x(l)$. If the transmission fails, the file size remains unchanged and therefore the next fragment remains $x(l)$, until the packet is successfully transmitted. Recall that F is the df of A_i. The expected time it takes to successfully transmit a fragment is $(x(l)+\phi)/\overline{F}(x(l)+\phi)$, the cost per trial multiplied by the expectation of the number of trials, which is geometrically distributed with parameter $F(x(l)+\phi)$. This implies that if we let $J(l) := \mathbb{E}\left[T(l)\right]$ denote the expected completion time when the file size is l under a generic Markov policy $x(l)$, then

$$J(l) \quad = \quad J(l - x(l)) + \frac{x(l)+\phi}{\overline{F}(x(l)+\phi)}.$$

Given any Markov policy $x(l)$, consider the sequence of fragments $x_1, x_2, \ldots$, generated from an initial file size L, defined recursively as:

$$x_1 := x(L); \ x_{i+1} := x(L - x_i), i \geq 1$$

such that $\sum_k x_k = L$. Define the expected time to successfully transmit a segment of size x as

$$h(x) \quad = \quad \frac{x+\phi}{\overline{F}(x+\phi)}. \quad (8)$$

The expected completion time is thus

$$J(L) \quad = \quad \sum_k h(x_k).$$

Since $h(x) \geq h(0) > \phi > 0$ for all $x \geq 0$, an optimal policy must only have finitely many terms in $J(L)$. Let $J^*(L)$ denote the (minimum) expected completion time under an optimal policy x^*.

Consider the following optimization problem:

$$\begin{aligned} H^* \ := \ \min_K \ \min_{y_1,\ldots,y_K} \quad & \sum_{k=1}^{K} h(y_k) && (9a) \\ \text{subject to} \quad & \sum_{k=1}^{K} y_k = L && (9b) \\ & y_k > 0, \quad k = 1, \ldots, K && (9c) \\ & K = 1, 2, \ldots. && (9d) \end{aligned}$$

We now argue that, given $L > 0$, the sequence of fragment sizes $x^* := (x_1^*, x_2^*, \ldots, x_{K^*}^*)$ generated by a Markov policy

$x^*(l)$ minimizes the expected completion time $\mathbb{E}[T(L)]$ if and only if (K^*, x^*) is a minimizer of (9a)–(9d). We can thus focus on solving (9a)–(9d). Indeed, we will show that under Assumption A1, (9a)–(9d) has a unique solution with $x_i^* = x^*$ for all i, implying that the optimal policy divides the file into equal sized fragments. [1]

Now, any finite sequence $(x_1, x_2, \ldots, x_K)$ with $\sum_k x_k = L$, $x_k > 0$ is a feasible solution of (9a)–(9d). Hence, $H^* \leq J^*(L)$. Conversely, given any minimizer (K^*, y^*) of (9a)–(9d), we will exhibit a Markov policy $x(l)$ that generates the sequence of fragment sizes that coincide with the given $y^* = (y_1^*, \ldots, y_{K^*}^*)$. This implies the minimum expected completion time satisfies $J^*(L) \leq H^*$. Hence, $J^*(L) = H^*$.

Parametrize the optimization problem (9a)–(9d) by the file size in (9b), and write any minimizer as $(K^*(l), y^*(l))$ when the file size is l. Consider the Markov policy $x(l)$ that solves (9a)–(9d) with file size l and selects the segment size $x(l) = y_1^*(l)$, i.e., the policy uses the first element of the solution $y^*(l)$ as the segment size when the remaining file size is l. The next segment size under policy $x(l)$ therefore comes from the solution of (9a)–(9d) with file size $l - x(l))$, i.e., $x(l - x(l)) = y_1^*(l - y_1^*(l))$. But $y_1^*(l - y_1^*(l))$ must be (equal to) the second element in the original solution, i.e., $y_1^*(l - y_1^*(l)) = y_2^*(l)$, for otherwise, $y^*(l)$ could not have been a minimizer. This implies by induction that the Markov policy $x(l)$ generates the sequence of fragment sizes from L that coincides with (K^*, y^*).

The main result of this section is the following theorem that says that the optimal policy creates equal sized fragments. The optimal fragment size depends on the file size.

$$g(x) = \frac{x + \phi}{x\overline{F}(x + \phi)} \tag{10}$$

and

$$a = \arg\min_x g(x), \qquad x \in \mathbb{R}_+ \tag{11}$$

Note that $g(x) = h(x)/x$ where $h(x)$ is the expected cost (time) to successfully transmit a segment of size x defined in (8). Hence we can interpret $g(x)$ as the per-bit cost for a fragment of size x, and a as the fragment size that minimizes the per-bit cost. It will become clear below that the optimal fragment size x^* is close to a and the minimum cost $J^*(L)$ is close to $Lg(a)$, under the following assumption:

A1: The density function $F' =: f$ exists. Moreover, the failure rate $\lambda(x) := f(x)/\overline{F}(x)$ is continuous and non-decreasing.[2]

Theorem 12 (Optimal fragmentation)**.** *Under assumption A1, for any $L > 0$, minimizers (K^*, x^*) of (9) is given by:*

1) *K^* equals $\lfloor L/a \rfloor$ or $\lceil L/a \rceil$ whichever produces a smaller value of $g(L/K^*)$.*
2) *$x_k^* = L/K^*$ for $k = 1, \ldots, K^*$.*

[1] We abuse notation and use x to denote a fragmentation policy, a vector of fragment sizes, or a scalar representing a constant fragment size, depending on the context; x^* denotes these quantities under an optimal policy.

[2] If $f(x) = \overline{F}(x) = 0$, define $\lambda(x) = \infty$.

Therefore, the optimal policy divides the file into K^* fragments of equal size. Each fragment is (re)submitted to the channel until the transmission is successful.

Proof of Theorem 12: We will first prove that, given any K, the minimizer x^* of the inner minimization exists, is unique, and $x_k^* = L/K$ for all k. We then prove that the optimal K^* is as stated in the theorem.

Given any integer $K > 0$, by (8), the KKT condition [13] for the inner optimization problem in (9a) implies that the optimum $x^* = (x_1^*, \ldots, x_K^*)$ satisfies, for all $k = 1, \ldots, K$,

$$\frac{dh(y_k)}{dy_k} = \frac{1}{\overline{F}(y_k + \phi)} + (y_k + \phi)\frac{f(y_k + \phi)}{\left(\overline{F}(y_k + \phi)\right)^2} = \lambda \tag{12}$$

By assumption A1, $\lambda(x) = f(x)/\overline{F}(x)$ is non-decreasing. Moreover $1/\overline{F}(x)$ is non-decreasing, and $x/\overline{F}(x)$ is strictly increasing. Therefore $h'(x)$ is strictly increasing, which is equivalent to $h(x)$ being strictly convex. Thus the inner minimization problem is strictly convex and the KKT condition is also sufficient. A unique solution $x^* = (x_1^*, \ldots, x_K^*)$ exists. Moreover, since all x_k^* are uniquely determined by (12), they are the same and hence $x_k^* = L/K$ for all k.

This reduces the minimization (9) to:

$$\min_K \qquad K\frac{L/K + \phi}{\overline{F}(L/K + \phi)} \;=\; L\frac{L/K + \phi}{L/K\overline{F}(L/K + \phi)}$$

Since L is constant, this is equivalent to solving

$$x^* = \arg\min_x g(x), \qquad x = \left\{L, \frac{L}{2}, \frac{L}{3}, \ldots\right\} \tag{13}$$

where g is defined in (10). The derivative of $g(x)$ is

$$\frac{dg(x)}{dx} = \frac{(x^2 + \phi x)f(x + \phi) - \phi\overline{F}(x + \phi)}{\left(x\overline{F}(x + \phi)\right)^2}$$

Since $\lambda(x) = f(x)/\overline{F}(x)$ is continuous by assumption, and since $\lim_{x\to 0} g(x) = \infty$ and $\lim_{x\to\infty} g(x) = \infty$, an optimal $x^* \in \{L, L/2, L/3, \ldots\}$ and hence optimal K^* exists. Moreover, any unconstrained minimum a of $g(x)$ must also be an extremum. Thus, setting $g'(x) = 0$ yields

$$\xi(x) \;:=\; \frac{f(x + \phi)}{\overline{F}(x + \phi)} \cdot \frac{x(x + \phi)}{\phi} \;=\; 1$$

Since $f(x + \phi)/\overline{F}(x + \phi)$ is non-decreasing, $x(x + \phi)/\phi$ is strictly increasing, $\xi(0) = 0$, $\lim_{x\to\infty} \xi(x) = \infty$, and $f(x)$ is continuous, it follows that the equation $\xi(x) = 1$ will have a unique solution, which is the unique minimizer a of $g(x)$ defined in (11). Moreover, it implies that $g(x)$ is unimodal. This means that x^* equal to $\lfloor L/a \rfloor$ or $\lceil L/a \rceil$, whichever produces a smaller $g(x)$ value. ∎

Note that since $g(0) = \infty$, the theorem implies that $K^* = 1$ if $L \leq a$. [14] provides a useful sufficient condition for Assumption A1: if f is log-concave, so is F. Since F is log-concave if and only if its failure rate is non-decreasing, a log-concave f satisfies A1. This is useful when F is hard to determine, e.g., for the Gaussian distribution.

The result of Theorem 12 applies to two failure models described in [7][pp. 131] - a model for satellite communication wherein A_i is exponentially distributed and a model for terrestrial communication, wherein A_i has a uniform distribution.

We now show that, when L is large, the unique optimal fragment size x^* is close to a; indeed, x^* approaches a as L increases.

Theorem 13. *Suppose $L > a$. Under assumption A1, the optimal fragment size $x^*(L)$ satisfies:*

1) $a/2 < x^*(L) \leq 2a$.
2) $a/(1 + a/L) < x^*(L) \leq a/(1 - a/L)$.

Proof: We know that for some integer K:

$$\frac{L}{K+1} \leq a < \frac{L}{K} \tag{14}$$

and

$$x^* = \frac{L}{K} \quad \text{or} \quad x^* = \frac{L}{K+1}$$

In the first case, $x^*K/(K+1) \leq a < x^*$ implying $x^*/2 \leq a < x^*$, i.e., $a < x^* \leq 2a$. In the second case, $x^* \leq a < x^*(K+1)/K \leq 2x^*$ implying $a/2 < x^* \leq a$. Combining yields $a/2 < x^* \leq 2a$.

From (14) we get

$$\frac{L}{a} - 1 \leq K < \frac{L}{a}$$

implying

$$a < \frac{L}{K} \leq \frac{a}{1 - a/L} \qquad \text{and} \qquad \frac{a}{1 + a/L} < \frac{L}{K+1} \leq a$$

Hence

$$\frac{a}{1 + a/L} < x^* \leq \frac{a}{1 - a/L}$$

■

This admits the following useful corollary.

Corollary 14.

$$\lim_{L \to \infty} x^*(L) = a.$$

B. Simple blind policy $x(l) = \min\{a, l\}$

The optimal fragmentation policy in Theorem 12 depends on the file size L. Consider the L-independent blind policy $x(l) = \min\{a, l\}$ where the fragment size a, given by (11), is always used until the remaining file size drops below a when it is transmitted in a single packet. We will abuse notation and use a to denote both this blind policy and the fragment size under this policy. Let $J^a(L)$ denote the expected file completion time under policy a when the file size is L. Recall that $J^*(L)$ denotes the minimum expected completion time. From Corollary 14, we know that policy a is asymptotically optimal, i.e., $x^*(L) \to a$. Hence we would expect $J^a(L)$ and $J^*(L)$ to be close for large L. The following result bounds their distance by an L-independent constant for any L.

Theorem 15. *Under Assumption A1, for any $L > 0$,*

$$\begin{aligned} 0 \;\leq\; J^*(L) - Lg^* \;&\leq\; h(a) \\ J^a(L) - J^*(L) \;&\leq\; h(a) \end{aligned}$$

where $h(x)$ is defined in (8) and $g^* := g(a)$ is defined by (10) *and* (11).

Proof: If $L = ka$ for some integer k, the proof of Theorem 12 shows that the policy a is optimal, in which case $J^a(L) = J^*(L)$. Suppose then that $ka < L < (k+1)a$ for some integer k. Clearly, $J^a(L) = kh(a) + h(L - ka)$. Since h is monotone, we have

$$kh(a) \leq J^a(L) \leq (k+1)h(a) \tag{15}$$

Since $J^*(L)$ is monotone in L, we have

$$kh(a) = J^*(ka) \leq J^*(L) \leq J^*((k+1)a) = (k+1)h(a) \tag{16}$$

Combining (15) and (16), we get that $J^a(L) - J^*(L) \leq h(a)$. This proves the sub-optimality bound. Moreover, (16) also implies $Lg^* \leq J^*(L) \leq Lg^* + h(a)$, as desired. ■

We make the following remarks:

1) Under both the optimal policy x^* and the blind policy a, the expected completion time grows (roughly) linearly in the file size, the approximating proportionality constant being the minimum per-bit cost $g(a)$.
2) The sub-optimality in expected completion time under the blind policy a is bounded by a constant independent of the file size.

C. Tail asymptotics under policies x^* and a

Denote by $T^*(L)$ and $T^a(L)$ respectively the completion times under the policies x^* and a.

Theorem 16. 1) *If L is light-tailed, then $T^*(L)$ and $T^a(L)$ are light-tailed.*
2) *If $L \in \mathcal{RV}(\alpha)$, then*

$$P\left(T^*(L) > t\right) \sim P\left(T^a(L) > t\right) \sim P\left(L > \frac{t}{g(a)}\right)$$

Since the blind policy a belongs to the class of constant fragmentation policies (see Section III), the tail asymptotics of $T^a(L)$ stated in the theorem follows from Theorem 9. Theorem 13 implies that the optimal policy x^* is a bounded fragmentation policy (see Section III). It follows then from Theorem 8 that $T^*(L)$ is LT if L is LT. However, the exact tail asymptotics of $T^*(L)$ when $L \in \mathcal{RV}(\alpha)$ claimed above requires a separate proof, which we omit due to space limitation.

V. Robustness to failure process

Although the blind policy of Section IV-B does not require knowledge of the file size L, it assumes knowledge of the statistics of the failure process $(A_n, n \geq 1)$. In this section, we derive bounds on the penalty for applying either the optimal policy x^* or blind policy a of Section IV designed for a failure distribution $\hat{F}$, when the actual distribution is F. Variables

with a hat will be used to denote quantities defined with respect to $\hat{F}$, e.g., $\hat{a}$ and $\hat{x}^*$ are the the blind and optimal policy, respectively, for the design distribution $\hat{F}$, while a and x^* are those for the true distribution F. Further, let $g^* := g(a) = \min_x g(x)$ where g is defined in (10).

We will compare the expected cost $J^{\hat{a}}(L)$ under F of the blind policy $\hat{a}$ designed for $\hat{F}$, and the expected cost $J^{\hat{x}^*}(L)$ under F of the policy $\hat{x}^*$ optimal for $\hat{F}$, with the true minimum cost $J^*(L)$. The following result specifies the cost increment in terms of the per-bit cost function g defined in (10).

Theorem 17. *Under assumption A1*

$$\begin{aligned}\lim_{L\to\infty} \frac{J^{\hat{a}}(L) - J^*(L)}{L} &= g(\hat{a}) - g^* \\ \lim_{L\to\infty} \frac{J^{\hat{x}^*}(L) - J^*(L)}{L} &= g(\hat{a}) - g^*\end{aligned}$$

Proof: To establish the first limit, note that for any constant fragment size x,

$$J^x(L) = \left\lfloor \frac{L}{x} \right\rfloor x g(x) + x' g(x'),$$

where $x' = L - \lfloor L/x \rfloor\, x \in [0, x)$. Since $x' g(x') = h(x')$, and $h(\cdot)$ is non-decreasing, this implies

$$|J^x(L) - Lg(x)| \;<\; h(x) \tag{17}$$

We also have $Lg^* \le J^*(L) \le Lg^* + h(a)$ from Theorem 15. Setting $x = \hat{a}$ in (17) then gives

$$J^{\hat{a}}(L) - J^*(L) \;=\; L\,(g(\hat{a}) - g^*) + \alpha(L) h(\max(a, \hat{a}))$$

for some $\alpha : \mathbb{R}_+ \to (-1, 1)$. Dividing the inequality by L and taking the limit as $L \to \infty$ gives the result.

The second inequality follows by setting $x = \hat{x}^*$ in (17) and following the same argument, noting that $\hat{x}^* \to \hat{a}$ and g is continuous. ∎

We make two remarks. First, without modeling error, $\hat{F} = F$, Theorem 15 implies that the per-bit cost penalty approaches zero as L increases. With modeling error, this penalty approaches $g(\hat{a}) - g^*$ which has the intuitive interpretation that the per-bit cost over the entire file approaches the per-bit cost over a packet. Second, an immediate corollary of Theorem 17 is that the overall per-bit costs of policies $\hat{a}$ and $\hat{x}^*$ are asymptotically the same, i.e.

$$\lim_{L\to\infty} \frac{J^{\hat{a}} - J^{\hat{x}^*}}{L} \;=\; 0 \tag{18}$$

which is also intuitive given $\hat{x}^* \to \hat{a}$.

The limit $g(\hat{a}) - g^*$ in Theorem 17 implies a bound on the per-bit cost penalty in terms of the error bound between the design distribution $\hat{F}$ and the true distribution F. Specifically, suppose the tail distributions satisfy

$$1 - F(x) = (1 - \hat{F}(x))(1 + \Delta(x)) \tag{19a}$$

where

$$-\Delta_{\min} \le \Delta(x) \le \Delta_{\max} \tag{19b}$$

for some $\Delta_{\min}$ and $\Delta_{\max}$. In that case, the cost penalty can be quantified in terms of the known quantities $\hat{g}^* := \hat{g}(\hat{a}) = \min_x \hat{g}(x)$, $\Delta_{\min}$ and $\Delta_{\max}$.

Theorem 18. *Under assumption A1*

$$\lim_{L\to\infty} \frac{J^{\hat{a}}(L) - J^*(L)}{L} \le \frac{\Delta_{\max} + \Delta_{\min}}{(1 + \Delta_{\max})(1 - \Delta_{\min})}\, \hat{g}^*$$

Proof: By Theorem 17, it suffices to show that the right hand side is at least $g(\hat{a}) - g^*$. By insertion of equation (19) into equation (10) we see that

$$\frac{\hat{g}(x)}{1 + \Delta_{\max}} \le g(x) \le \frac{\hat{g}(x)}{1 - \Delta_{\min}} \tag{20}$$

Since equation (20) holds for a, we get

$$\frac{\hat{g}^*}{1 + \Delta_{\max}} \le \frac{\hat{g}(a)}{1 + \Delta_{\max}} \le g^* \tag{21}$$

and for $\hat{a}$, from which we get

$$g(\hat{a}) \le \frac{\hat{g}^*}{1 - \Delta_{\min}} \tag{22}$$

Combining inequalities (21) and (22), we get

$$\begin{aligned} g(\hat{a}) - g^* &\le \frac{\hat{g}^*}{1 - \Delta_{\min}} - \frac{\hat{g}^*}{1 + \Delta_{\max}} \\ &= \frac{\Delta_{\max} + \Delta_{\min}}{(1 + \Delta_{\max})(1 - \Delta_{\min})}\, \hat{g}^* \end{aligned}$$

as required. ∎

Corollary 19. *If $\Delta_{\min} = \Delta_{\max}$, under assumption A1,*

$$\lim_{L\to\infty} \frac{J^{\hat{a}}(L) - J^*(L)}{L} \le \frac{2\Delta_{\max}}{1 - (\Delta_{\max})^2}\, \hat{g}^*$$

VI. Related work

Optimal file fragmentation that maximizes throughput is considered in the early work of [7][pp. 131] which, using a specific failure model and assuming equal-sized packets, derives an expression for optimal fragment size. This result can be considered as a special case of ours on optimal fragmentation.

Besides the recent work mentioned in Section I, a different application that is mathematically related to our problem is checkpointing. In a checkpointing problem, a failure can occur during the execution of a program (job) and when that happens, recovery is initiated after which the program must be restarted. Checkpoints can be inserted so that program execution can be restarted from the last checkpoint before the failure instead of from the beginning. There is a sizeable literature on checkpointing; see [15] for an early survey and also references in, e.g., [16]. The model that is closest to ours is that (Model 1) in [17] where the authors study optimal checkpointing to minimize expected completion time and derive the Laplace transform of the completion time. They show that, when the inter-failure times are exponential or uniform, equally spaced checkpoints are optimal and provide an expression for the optimal size of program fragments between checkpoints. We prove the same result in Section

IV-A that fragmentation of the file into equal-sized fragments minimizes expected completion time as long as the failure rate is nondecreasing (both exponential and uniform distributions satisfy this condition). Even though a Laplace transform of the completion time can be used to deduce tail behavior, we provide a simpler characterization of the tail behavior of completion time through direct analysis on its distribution. A similar model to [17] is considered in [18] which focuses on *online* checkpointing algorithms that dynamically estimate checkpointing and recovery costs and place a checkpoint either when the cost is small or when a long time has elapsed since the last checkpoint. The paper shows that the performance of this strategy is close to the optimal offline algorithm that knows all costs in advance. Equally spaced checkpointing has also been analyzed in various other models. For instance, [19] derives an expression for the optimal size of checkpoint intervals when inter-failure times are exponential, and shows that the expected completion time is exponential to the job size L without checkpointing and (roughly) linear in L with checkpointing. In [16], the authors consider a program of infinite length and derive the optimal checkpointing frequency that minimizes the expected cost of setting up checkpoints and the work that is wasted when a failure occurs. They formulate the optimization as a variational problem and use the Euler-Lagrange equation to show that the optimal frequency is (roughly) proportional to the square root of the failure rate. In the special case when the inter-failure times are exponential, equally spaced checkpoints are optimal. In [20] the authors formulate the problem of minimizing expected completion time as a continuous-state continuous-time dynamic program, and show how it can be numerically solved by discretizing state and time and applying value or policy iteration methods for finite-state problems. In [21] the authors consider an infinite-program model where the system goes from the normal state into the repair state or the checkpointing state according to independent Poisson processes. The main result is the derivation of the Laplace transform of the completion time distribution and its first two moments. Even though the models in the checkpointing literature are slightly different, some of the insights and techniques are applicable to the models here and that in [1]–[4]. See [8] for a set of results parallel to those reported in this paper for a *job* fragmentation model that is closer to the models described in this section.

VII. Conclusion

It has been discovered that file completion time on an unreliable channel can be heavy-tailed even when file size distribution is not. To mitigate, we show that independent or bounded file fragmentation guarantees light-tailed file completion time as long as the file size is light-tailed. When the file size is heavy-tailed (specifically, regularly varying), the completion time is as heavy-tailed as the file size (in the degree sense). Finally, seeking to minimize the expected completion time, we derive the optimal fragmentation policy as well as a simple suboptimal blind fragmentation policy. We also characterize the robustness of these policies to error in modeling the failure process.

Acknowledgment

We thank Adam Wierman, Lijun Chen and Mani Chandy for helpful discussions. We acknowledge support of ARO through MURI Grant W911NF-08-1-0233, NSF through the NetSE grant, the Caltech Lee Center for Advanced Networking, and Australian Research Council grant DP0985322.

References

[1] R. Sheahan, L. Lipsky, P. M. Fiorini, and S. Asmussen, "On the completion time distribution for tasks that must restart from the beginning if a failure occurs," *ACM SIGMETRICS Performance Evaluation Review*, vol. 34, no. 3, pp. 24–26, December 2006.

[2] S. Asmussen, P. Fiorini, L. Lipsky, T. Rolski, and R. Sheahan, "Asymptotic behavior of total times for jobs that must start over if a failure occurs," *Mathematics of Operations Research*, vol. 33, no. 4, pp. 932–944, 2008.

[3] P. R. Jelenković and J. Tan, "Characterizing heavy-tailed distributions induced by retransmissions," Department of Electrical Engineering, Columbia University, New York, NY, Technical Report EE2007-09-07, September 2007.

[4] ——, "Can retransmissions of superexponential documents cause subexponential delays?" in *Proceedings of IEEE INFOCOM'07*, 2007.

[5] ——, "Dynamic packet fragmentation for wireless channels with failures," in *MobiHoc '08: Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing*. New York, NY, USA: ACM, 2008.

[6] J. Tan, W. Wei, B. Jiang, N. Shroff, and D. Towsley, "Can multipath mitigate power law delays?" 2009, submitted for publication.

[7] M. Schwartz, *Telecommunication networks: protocols, modeling and analysis*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1986.

[8] J. Nair and S. H. Low, "Optimal job fragmentation," *SIGMETRICS Perform. Eval. Rev.*, vol. 37, no. 2, pp. 21–23, 2009.

[9] N. H. Bingham, C. M. Goldie, and J. L. Teugels, *Regular Variation (Encyclopedia of Mathematics and its Applications)*. Cambridge University Press, 1987.

[10] S. Resnick, *Heavy-Tail Phenomena:Probabilistic and Statistical Modeling*. Springer, 2007.

[11] K. Sigman, "Appendix: A primer on heavy-tailed distributions," *Queueing Syst. Theory Appl.*, vol. 33, no. 1-3, pp. 261–275, 1999.

[12] F. Guillemin, P. Robert, and B. Zwart, "Tail asymptotics for processor-sharing queues," *Advances in Applied Probability*, vol. 36, no. 2, pp. 525–543, 2004.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[14] M. Bagnoli and T. Bergstrom, "Log-concave probability and its applications," Department of Economics, UC Santa Barbara, University of California at Santa Barbara, Economics Working Paper Series, Jan. 2004. [Online]. Available: http://ideas.repec.org/p/cdl/ucsbec/1989d.html

[15] K. Chandy, "A survey of analytic models for rollback and recovery strategies," *Computer*, vol. 8, no. 5, pp. 40–47, 1975.

[16] Y. Ling, J. Mi, and X. Lin, "A variational calculus approach to optimal checkpoint placement," *IEEE Trans. on Computers*, vol. 50, no. 7, pp. 699–708, 2001.

[17] G. C. Jr and N. Gilbert, "Optimal strategies for scheduling checkpoints and preventive maintenance," *IEEE Trans. on Reliability*, vol. 39, no. 1, pp. 9–18, April 1990.

[18] A. Ziv and J. Bruck, "An on-line algorithm for checkpoint placement," *IEEE Trans. on Computers*, vol. 46, no. 9, pp. 976–985, 1997.

[19] A. Duda, "The effects of checkpointing on program execution time," *Inf. Process. Lett.*, vol. 16, no. 5, pp. 221–229, 1983.

[20] P. Lécuyer and J. Malenfant, "Computing optimal checkpointing strategies for rollback and recovery systems," *IEEE Trans. Computers*, vol. C-37, no. 4, pp. 491–496, April 1988.

[21] V. Kulkarni, V. Nicola, and K. S. Trivedi, "Effects of checkpointing and queueing on program performance," *Communications in Statistics – Stochastic Models*, vol. 6, no. 4, pp. 615–648, 1990.