A Particle Process Underlying SSD Storage Structures

[Extended Abstract]

Edward G. Coffman
Electrical Engineering and Computer Science
Columbia University
New York, NY 10025
coffman@cs.columbia.edu

Petar Momčilović
Industrial and Systems Engineering
University of Florida
Gainesville, FL 32611
petar@ise.ufl.edu

ABSTRACT
We introduce a particle process that models the evolution of page configurations in solid-state-drive (SSD) storage devices. These devices use integrated circuitry as memory to store data. Typically, pages (units of storage) are organized into blocks of a given size. Three operations are permitted: write, read, and clean. Rewrites are not allowed, i.e., a page has to be “cleaned” before the write operation can be repeated. While the read and write operations are permitted on individual pages, the clean operation can be executed on whole blocks only. Analysis of our particle process captures a key tradeoff in the operation of SSD’s.

1. INTRODUCTION
Partition $\mathbb{Z}^+$ into consecutive size-$k$ blocks $\{(j-1)k+1, \ldots, jk\}$, $j = 1, 2, \ldots$. We define a particle process with parameters $\lambda > 0$, $k \geq 1$ and $C_k \geq 0$ as follows. The $i$-th particle in a stationary, rate-$\lambda$ arrival process is placed at integer $i$ and remains active there for a holding time sampled from a given distribution independently of all other particle holding times. When its holding time expires a particle becomes inactive until it is removed. An inactive particle in the $j$-th block is removed when, and only when, inactive particles exist at all other integers in the $j$-th block; the time to remove the $k$ inactive particles in a block is a constant $C_k$.

We say that a block is active if it has at least one active particle; otherwise it is either inactive, i.e., fully occupied by inactive particles, or available, i.e., no particles have yet been placed in the block, or all particles placed in the block so far have been removed. Our interest focuses on the steady-state of the processes $N_k(t)$ and $M_k(t)$ giving, at time $t$, the number of active blocks and inactive blocks, respectively.

The connection with a baseline stochastic model of SSD storage technology is implicit in the following changes in terminology: Replace particle by “file,” integer $i$ by “page” $i$, and the notion of removing particles from a block by the notion of “cleaning” a block. Then the particle process models an infinite SSD memory cell with $k$-page blocks, and with file sizes equal to the page size. The remainder of the paper stays with the natural SSD terminology as it keeps the application in view at no cost.

SSD performance studies are an increasingly important, fast growing research area. Space constraints prevent us from covering the literature in full; for a much more comprehensive coverage of the literature, we refer the reader to two, just published papers [1, 3]. Each of these papers studies interesting garbage-collection optimization problems, and both apply mean-field approximation techniques. Garbage collection refers to the cleaning of blocks having a relatively large fraction of inactive files (garbage), a process that must be preceded by copying to clean pages in other blocks the relatively few still-active files. Garbage collection reclaim clean storage space, and makes for more efficient utilization of storage, but it comes with non-trivial costs: Cleaning times are relatively long, and blocks have finite lifetimes in the sense that there is a limit on the number of cleaning operations that a block can sustain – a characteristic feature of SSD storage.

2. THE PROBABILITY MODEL
We use the following notation throughout. For $f, g : \mathbb{R} \to \mathbb{R}$, $f(x) \sim g(x)$, as $x \to \infty$, stands for $f(x)/g(x) \to 1$, as $x \to \infty$. Symbols $\wedge$ and $\vee$ denote the maximum and minimum operators, respectively. Given a nondecreasing function $f$ on $\mathbb{R}$, $f^{-\infty}$ is the (left-continuous) inverse of $f$: $f^{-\infty}(x) = \inf\{y : f(y) \geq x\}$.

The holding times requested by files are i.i.d. with the distribution $F(\bar{F} = 1 - F)$, such that $F(0) = 0$ and $\mu^{-1} = \int_0^\infty F(x) dx < \infty$. However, the actual holding times have an “archiving” parameter $L_k > 0$. A file is made inactive after $L_k$ time units in the system, if the requested holding time is larger than $L_k$. The actual holding time distribution is then given by $F_\ast(x) = 1 - F(x) = 1 - \bar{F}(x)_{[x < L_k]}$. The assumption is that, if a file’s holding time is found to exceed $L_k$, it is moved to archival storage, which is in a region separate from that being modeled here. We return to models with $L_k < \infty$ in a little more detail in Section 3.

Let $A(t)$ be the number of arrivals in the time interval $[0, t]$. Process $\{A(t), t \geq 0\}$ is stationary with rate $\lambda$; define $\rho = \lambda / \mu$. Since the $i$th arriving file is stored in the $i$th page, and since blocks are cleaned only when all files in a block are inactive, then the total number of active blocks at time $t$ is given by

$$N_k(t) = \sum_{i=1}^{a_k(t)} Y_{k,i}(t),$$

where $a_k(t) = \lceil A(t)/k \rceil$ and $\{Y_{k,i}(t)\}$, are Bernoulli random variables; $Y_{k,i}(t) = 1$ corresponds to the case when the block $a_k(t) + 1 - i$ is active (contains at least one active file), while $Y_{k,i}(t) = 0$ otherwise. Likewise, let $Z_{k,i}(t) = 1$ if block

MAMA 2013 Pittsburg, PA USA
Copyright is held by author/owner(s).
Let $X(t) = 1$ if page $A(t) + 1 - i$ is active, i.e., has an active file, and let $X_i(t) = 0$ otherwise. Then

$$Y_{k,i}(t) = \int_{j=a_k(t)-i+1}^{A(t)} X_j(t);$$

$$\mathbb{P}[X_i(t) = 1] = \tilde{F}(t - \inf\{u > 0 : A(u) = i\}),$$

for $i = 1, 2, \ldots, A(t); i = 1$ and $i = A(t)$ correspond to the last and the first files to arrive, respectively.

Let $X, Y_{k,i}, Z_{k,i}, N_k$ and $M_k$ be random variables equal in distribution to $X_i(t), Y_{k,i}(t), Z_{k,i}(t), N_k(t)$ and $M_k(t)$ in steady-state (as $t \to \infty$). Then, we have

$$Y_{k,i} = \int_{j=U_k + (i-2)k+1}^{U_k + (i-1)k+1} X_j,$$

where $U_k$ is a uniform random variable on $\{1, 2, \ldots, k\}$. Let $\{T_i\}$ be a sequence of arrival times. Then, $N_k = \sum_{i=1}^{N_k} Y_{k,i}$,

$$\mathbb{P}[Y_{k,i} = 0] = \mathbb{P} \left[ \bigcup_{j=(U_k + (i-2)k+1)}^{U_k + (i-1)k+1} X_j = 0 \right] = \mathbb{E} \prod_{j=(U_k + (i-2)k+1)}^{U_k + (i-1)k+1} F_T(T_j), \quad (1)$$

and

$$\mathbb{P}[Z_{k,i} = 1] = \mathbb{E} \prod_{j=(U_k + (i-2)k+1)}^{U_k + (i-1)k+1} F(T_j) - \mathbb{E} \prod_{j=(U_k + (i-2)k+1)}^{U_k + (i-1)k+1} F(T_j - C_k).$$

In the case of Poisson arrivals, $k = 1$ and $L_k = \infty$, $N_k$ is equal in distribution to equal in the steady-state number of customers in an $M/G/\infty$ process. $N_k$ is known to be Poisson with mean $\rho$ [4, p. 79] in this case. This offers a check of (1) which gives:

$$\mathbb{E}(z + 1)^{N_k} = \mathbb{E} \prod_{i=1}^{N_k} \left( 1 + z \tilde{F}(T_i) \right) = e^{z \lambda \int_0^{L_k} F(x) dx} = e^{z \rho},$$

where the second equality is due to the fact that $T_i$’s are arrival points in a Poisson process with intensity $\lambda$.

The following is our main result.

**Theorem 1.** Consider our particle process with the following scalings for $k$.

- **(Small $k$).** Let $k$ be fixed. Then $\mathbb{E} \sum_{i=1}^{N_k} \mathbb{E} \lambda C_k/k$ and

$$\mathbb{E} N_k \sim \frac{\lambda}{k} \int_0^{L_k} \left( 1 - F^k(x) \right) dx,$$

as $\lambda \to \infty$.

- **(Moderate $k$).** Suppose $k \to \infty$ and $k = o(\lambda)$, as $\lambda \to \infty$. Assume that the holding time distribution satisfies, as $x \to \infty$, $F(sx)/F(x) \to s^{-\alpha}$, for some $\alpha \in (1, \infty)$. Then, as $\lambda \to \infty$, $\mathbb{E} N_k \sim \lambda C_k/k$ and

$$\mathbb{E} N_k \sim \frac{\lambda}{k} \left( \tilde{F}^{-1} \right)^{\alpha} \left( k \left( \frac{2 - \alpha}{\alpha} \right) \tilde{L}^\alpha + \left( 1 - e^{-\tilde{L}^\alpha} \right) \tilde{L} \right),$$

where $\tilde{L}_k = L_k/(\tilde{F}^{-1})^\alpha(k)$ and $\Gamma(\cdot, \cdot)$ is the incomplete gamma function.

- **(Large $k$).** Suppose $\lambda/k \to c \in [0, \infty)$, as $\lambda \to \infty$. Let $U$ be uniform on $[0, 1]$. Then, as $\lambda \to \infty$,

$$\mathbb{E} N_k \sim \sum_{i=1}^{\infty} \left( 1 - e^{-\lambda \frac{k}{k(U+i+1)\cdot L_k}} F(x) dx \right)$$

and

$$\mathbb{E}[N_k + M_k] \sim \sum_{i=1}^{\infty} \left( 1 - e^{-\lambda \frac{k}{k(U+i+1)\cdot L_k}} F(x) dx \right).$$

Next, we consider two examples.

**Example 1.** Suppose $F(x) = x$, $x \in [0, 1]$ and $L_k = \infty$.

Then

$$\mathbb{E} N_k \sim \frac{\lambda}{k + 1},$$

as $\lambda \to \infty$, for small and moderate $k$ ($\alpha = \infty$ in this case), and

$$\mathbb{E} N_k \sim \frac{\lambda}{k} \left( \frac{4^k}{2^k} - 1 \right),$$

as $\lambda \to \infty$, for fixed $k$; $\mathbb{E} N_k \sim \lambda \sqrt{n/k}$, as $\lambda \to \infty$, for moderate $k$; and

$$\mathbb{E} N_k \sim \sum_{i=1}^{\infty} \left( 1 - e^{-\frac{k}{k(U+i+1)} \left( (1-x)^2 \right) dx} \right),$$

as $\lambda \to \infty$, for large $k$. In particular, if $k/\lambda \to 1$, then, $\mathbb{E} N_k \to 2$, as $\lambda \to \infty$.

**Example 2.** Suppose $F(x) = +1 + x)^2$, $x \geq 0$, and $L_k = \infty$.

Then

$$\mathbb{E} N_k \sim \frac{\lambda}{k} \left( \frac{4^k}{2^k} - 1 \right),$$

as $\lambda \to \infty$, for fixed $k$; $\mathbb{E} N_k \sim \lambda \sqrt{n/k}$, as $\lambda \to \infty$, for moderate $k$; and

$$\mathbb{E} N_k \sim \sum_{i=1}^{\infty} \left( 1 - e^{-\frac{k}{k(U+i+1)} \left( (1-x)^2 \right) dx} \right),$$

as $\lambda \to \infty$, for large $k$. In particular, if $k/\lambda^2 \to c \in (0, \infty)$, then, as $\lambda \to \infty$,

$$\mathbb{E} N_k \sim 1 + \sum_{i=1}^{\infty} \left( 1 - e^{-c(U+i+1)^2} \right).$$

**Sketch of the proof.** Here we outline an approximate analysis when $\lambda \to \infty$. In that case, $T_i \approx i/\lambda$. Multiple regimes for $k$ need to be considered.

**Fixed $k$.** In this case, $T_i \approx T_{ik}$ for $(U_k + (i-2)k)^+ < j \leq U_k + (i-1)k$ in (1). That is, one can assume that all files for a single block arrive simultaneously. Hence,

$$\mathbb{E} N_k \approx \sum_{i=1}^{\tilde{L}_k} \left( 1 - F^k(T_{ik}) \right) \approx \frac{\lambda}{k} \int_0^{L_k} \left( 1 - F^k(x) \right) dx,$$

where the upper limit in the summation is due to $\tilde{F}_c(L_k) = 0$. The total number of blocks that are either active or are undergoing cleaning can be obtained in a similar way:

$$\mathbb{E} \sum_{i=1}^{\tilde{L}_k} \left( 1 - F^k(x_{ik}) \right) \approx \frac{\lambda}{k} C_k + \mathbb{E} N_k.$$
Moderate \( k \). Here one can still assume that all files arrive to a block simultaneously (since \( k/\lambda \to 0 \)). However, due to \( k \to \infty \), \( F^k(T_{ik}) \) can be approximated as

\[
F^k(T_{ik}) = (1 - \tilde{F}(T_{ik}))^k \approx e^{-k\tilde{F}(T_{ik})}.
\]

This leads to the following:

\[
\mathbb{E}N_k \approx \sum_{i=1}^{\lambda L_k} (1 - e^{-k\tilde{F}(T_{ik})}) \approx \frac{\lambda}{k} \int_0^{L_k} (1 - e^{-kF(x)}) \, dx
\]

\[
\approx \frac{\lambda}{k} (\tilde{F}^{-1})^{-(k)} \left( \int_0^{\tilde{L}_k} (1 - e^{-x}) \, dx \right)
\]

\[
\approx \frac{\lambda}{k} (\tilde{F}^{-1})^{-(k)} \left[ \left( \frac{\alpha - 1}{\alpha} \right) \tilde{L}^{-\alpha} + (1 - e^{-\tilde{L}^{-\alpha}}) \tilde{L} \right],
\]

where we used a change of variables and \( k\tilde{F}(x) \approx \tilde{F}(x) (\tilde{F}^{-1})^{-(k)}(k) \approx x^{-\alpha} \).

As in the previous case, similar reasoning can be used to estimate \( \mathbb{E}M_k \):

\[
\mathbb{E}M_k + \mathbb{E}N_k \approx \sum_{i=1}^{\infty} \left( 1 - e^{-k\tilde{F}(x)} \right)
\]

\[
\approx \frac{\lambda}{k} \int_0^{\infty} (1 - e^{-k\tilde{F}(x)}) \, dx \approx \frac{\lambda}{k} C_k + \mathbb{E}N_k.
\]

Large \( k \). In this case, \( k \) and \( \lambda \) are proportional, and one has to account for the fact that blocks are filled up gradually. Consequently:

\[
\mathbb{E}N_k \approx \sum_{i=1}^{\infty} \left( 1 - e^{-\sum_{j=0}^{i-1} \tilde{F}(T_{ik})} \right)
\]

\[
\approx \sum_{i=1}^{\infty} \left( 1 - e^{-\frac{\lambda}{k} \int_0^{L_k} (1 - e^{-x}) \, dx} \right).
\]

This concludes the sketch of the proof. \( \square \)

Remark 1. An intermediate regime can be analyzed by exploiting extreme value theory [2, Ch. 1]. In particular, \( F \in D(\Lambda) \) if there exist normalizing constants \( a_n \geq 0, b_n \in \mathbb{R} \) such that \( F_0^n(a_nx + b_n) \to G(x) \), as \( n \to \infty \), where \( G \) is one of the three extreme value distributions (\( \Lambda(x) = \exp\{-e^{-x}\} \), \( \Phi_{\alpha}(x) = \exp\{-x^{-\alpha}\} \), and \( \Psi_{\alpha}(x) = \exp\{-(-x)^{-\alpha}\} \)). Then, for large \( k \) (and hence large \( \lambda \)), we have

\[
\mathbb{E}N_k \approx \sum_{i=1}^{\infty} \left( 1 - F^k(T_{ik}) \right) \approx \frac{\lambda}{k} \int_0^{\infty} \left( 1 - F^k(x) \right) \, dx
\]

\[
\approx \frac{\lambda a_k}{k} \int_{-b_k/a_k}^{\infty} \left( 1 - F^k(a_kx + b_k) \right) \, dx
\]

\[
\approx \frac{\lambda a_k}{k} \int_{-b_k/a_k}^{\infty} (1 - G(x)) \, dx.
\]

Depending on the particular form of \( G \), a more explicit expression can be obtained for large \( k \):

\[
\mathbb{E}N_k \approx \frac{\lambda}{k} \int_{-b_k/a_k}^{\infty} \left( 1 - e^{-e^{-x/k}} \right) \, dx, \quad F \in D(\Lambda)
\]

\[
\mathbb{E}N_k \approx \frac{\lambda}{k} \int_0^{\infty} \left( 1 - e^{-x^{a/k}} \right) \, dx, \quad F \in D(\Phi_{\alpha})
\]

\[
\mathbb{E}N_k \approx \frac{\lambda}{k} \int_{-b_k/\mu}^{\infty} \left( 1 - e^{-x} \right) \, dx, \quad F \in D(\Psi_{\alpha})
\]

where \( x_* = \sup\{ x : F(x) < 1 \}, x_*^k = x_* - (\tilde{F}^{-1})^{-(k)}(k) \) and \( \gamma \) is the Euler-Mascheroni constant. For example, if \( F(x) = 1 - e^{-\mu x}, x \geq 0 \), then \( F \in D(\Lambda) \) with \( a_k = 1/\mu \) and \( b_k = \log k/\mu \). This leads to \( \mathbb{E}N_k \sim \rho \log k/k \), which also follows from our theorem.

3. ON-GOING RESEARCH DIRECTIONS

Archiving: Consider a system where, upon spending \( L_k \) time units in the system, a file is copied to an archive. Then, it is appropriate to think of the archive as a memory system with the arrival rate \( \bar{\Lambda}(L_k) \) and the holding time distribution \( 1 - \tilde{F}(x + L_k)/\tilde{F}(L_k), x \geq 0 \). We require that \( L_k \) be such that \( \bar{\Lambda}(L_k) \to \infty \), i.e., the archive operates in the high-volume regime as well. Under this model, it is possible to optimize \( L_k \) in order to minimize the total number of blocks used in main storage and archival storage.

Example 3. Suppose we have \( F(x) = e^{-x}, \) moderate \( k \), and instantaneous cleanups (\( C_k = 0 \)). The total number of blocks in use is approximately \( \frac{1}{k} (L_k + e^{-L_k} \log k) \), and thus the optimal time to archive files is \( L_k \approx \log \log k \). In that case, on average \( \approx \frac{1}{k} \log \log k \) blocks are utilized instead of \( \approx k/2 \log k \) without archiving.

Active Storage: Let \( r_1(t) \) denote the smallest index \( i \) such that the \( i \)-th block is active, and let \( r_2(t) \) denote the index of the block where the latest arrival in \( [0,t] \) was placed. The span from \( r_1(t) \) to \( r_2(t) \) extends the notion of active pages and active blocks to active storage. The expected value of the span \( r_2(t) - r_1(t) + 1 \) of active storage in the limit of large \( t \) is of obvious interest.

4. REFERENCES


