for both FF and BF, where the $n/2$ is contributed by the expected total item size. However, we claim that this result can be obtained directly from Theorem 1 for the stochastic model; we intend to supply a rigorous proof of this intuitive property in the near future. The point of this remark is that proofs of results like (1) seem to be much easier using the fluid limit approach. Of course, the fluid limit does not give a result more precise than $o(n)$ (a long and arduous proof in [2] shows that expected wasted space in the fixed-input model is in fact $O(n^{2/3})$).

The proof of Theorem 1 begins by formulating the underlying Markov process in such a way that the fluid scaling makes sense. For this, we let $F_i(t)$ and $\hat{F}_i(t)$ be the arrival and departure processes for the type-i items; clearly, the state of the queue is uniquely recoverable from $(F(t), \hat{F}(t))$ for all $t \geq 0$, where $F = (F_1, \ldots, F_k)$ and $\hat{F} = (\hat{F}_1, \ldots, \hat{F}_{k-1})$. Next, introduce a superscript (n) to denote the fact that the initial state has n waiting in queue, and let $f = (f_1, \ldots, f_{k-1})$ and $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_{k-1})$, where $f_i = f_i(t)$ is the limit $n \to \infty$ of the scaled process $\frac{1}{n}(F_i^{(n)}(t), t \geq 0)$, and where $\hat{f}_i$ is defined similarly in terms of $\frac{1}{n}\hat{F}^{(n)}(t), t \geq 0)$.

The goal of the analysis is to show that there exists a $T > 0$ such that, with probability 1, the queue-size limit $q(t) := \sum_i f_i(t) - \hat{f}_i(t)$ 'drains' to 0 by time T and stays there; an appeal to well-known results then concludes the proof. To prove this fact for BF requires only a rather compact inductive proof that, whenever $\hat{f}_i(t) < f_i(t)$ for t sufficiently large, $\hat{f}_i(t)$ drifts towards $f_i(t)$ at a rate exceeding $1 - \rho > 0$. The key observation that makes the proof for FF easy is that the analysis is much simpler in terms of the function $\tau(t) := \min_i \tau_i(t)$, where $\tau_i(t) := \inf\{s < t : f_i(s) > \hat{f}_i(t)\}$. As can be seen, the empty-queue condition $q(t) = 0$ is $\tau_i(t) = t$ for all i. The proof shows that, whenever $\tau(t) < t$, there is a drift $\tau'(t) - 1 > 1/\rho - 1 > 0$ towards the empty-queue condition.

References

1. Coffman, E. G., Jr., D. S. Johnson, P. W. Shor, and R. R. Weber, "Bin Packing with Discrete Item Sizes, Part II: Tight Bounds on First Fit," *Random Structures and Algorithms*, **10**(1997), pp. 69–101.

2. Coffman, E. G., Jr. and A. L. Stolyar, "Bandwidth Packing," (in preparation).

3. Rybko, A. N. and A. L. Stolyar, "Ergodicity of Stochastic Processes Describing the Operation of Queueing Networks," *Problems of Information Transmission*, **28**(1992), pp. 199–220.

Fluid Limits, Bin Packing, and Stochastic Analysis of Algorithms

E. G. Coffman, Jr. and A. L. Stolyar
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974

For many years, fluid-limit techniques have provided useful asymptotics for evaluating the performance of computer and communication systems. Several years ago, Rybko and Stolyar [3] introduced fluid limits in the stability analysis of queueing network processes (see also Dai and Meyn [ ] for further results and key references), and within the last couple of years, the present authors demonstrated the potential of these techniques in the stochastic analysis of algorithms [3]. This short note gives a rough sketch of fluid limits and their application to the analysis of bin packing algorithms; the intent is to show the power of the fluid-limit approach and the promise it holds as a novel tool for the probabilistic analysis of algorithms.

The behavior approximated by fluid limits is long-run expected-value behavior. For example, let $(F(t),\ t \geq 0)$ be a rate-λ Poisson arrival (counting) process and consider the *fluid-limit scaling* $(\frac{1}{n}F(nt),\ t \geq 0)$. The (fluid) limit $n \to \infty$ of the scaled process is simply a straight line with slope λ and a position determined by a given initial condition. As illustrated below, in certain important cases, properties of a stochastic process can be inferred from a suitably defined, corresponding property of its fluid limit; the power of the approach then lies in the fact that arguments based on fluid limits are typically much simpler.

To illustrate this tool in action, consider the following discrete-time model of bin packing that arises in multimedia communication systems, computer storage allocation, and transmission along slotted communication channels [2]. Items with sizes in $(0, 1]$ arrive randomly at integer times and join a queue waiting to be packed into unit-capacity bins. At each integer time, a set of items summing to < 1 is drawn from the queue and packed in the next empty bin according to a given algorithm. At the end of the unit interval devoted to packing a bin, new item arrivals, if any, are admitted to the queue; the number of items arriving at time t has moments of all orders and is independent of the numbers arriving at earlier times. For simplicity, we discretize item sizes so that each falls in the set $\{1/k, \ldots, (k-1)/k\}$ for some given $k > 1$. Let $\lambda_i,\ 1 \leq i \leq k-1$, be the arrival rate of *type-i* items, i.e., those with size i/k.

Consider the First Fit (FF) and Best Fit (BF) packing algorithms, both scanning the entire queue at each integer time and greedily packing into the current bin each item encountered that fits in the space left over by items already packed. FF scans the queue in arrival order and BF scans the queue in non-increasing order of item size. Let Q have the stationary distribution of queue size, assuming that one exists. We proved recently, using a fluid-limit technique,

Theorem If the $\lambda_i,\ 1 \leq i \leq k-1$, are symmetric, i.e., $\lambda_i = \lambda_{k-i}$, then for $\mathrm{E}Q < \infty$ to hold, it is necessary and sufficient that $\rho := \sum_i \lambda_i(i/k) < 1$.

Remark. Although the fluid limit, as described, applies to the analysis of algorithms in a stochastic setting, the results can be carried over to the standard set-up of average-case analysis. In terms of our model, the latter analysis assumes that n i.i.d. items are waiting in queue, the arrival process is turned off, and the goal is to find the expected value of N_n, the time (number of bins) needed to empty the queue (pack the n items). We know that the expected wasted time (space) is $o(n)$ under both FF and BF and hence asymptotically negligible; in particular,

$$\mathrm{E}N_n = \frac{n}{2} + o(n)\ as\ n \to \infty \tag{1}$$