OPERATIONAL ANALYSIS

Peter J. Denning

Computer Science Department
Naval Postgraduate School
Monterey, CA 93943

v1 -- 5/31/04
v2 -- 7/27/04

In 1975, Jeff Buzen and I discovered we shared a concern about a fundamental problem we observed in the field of performance analysis. This article is about that concern, what we did about it, and a key role that Ken Sevick played in the outcome.

Dead Cows

My friend and teacher, Fernando Flores, likes to tell his business clients the story of Pasteur and the dead cows. In the 1870s the French sheep and cattle industry was being decimated by anthrax and France's economic position was in grave peril. At the time, the farmers were completely baffled by the mounting toll of dead cows and sheep; the cause, anthrax, had not yet been identified and named. Although a few researchers, including Pasteur, believed that a microbe might be the cause, the theory that diseases could be caused by microbes was so much at odds with conventional thinking that few physicians accepted it; Pasteur could not even persuade surgeons to wash their hands and use clean instruments. Finally, in 1882, Pasteur was challenged to "put up or shut up" by a French veterinarian. Pasteur vaccinated 25 of a group of 50 sheep with his anthrax vaccine; all 50 then received a lethal dose of anthrax. Every one of the vaccinated sheep survived without symptoms and every one of the unvaccinated sheep died within three days. Pasteur became a national hero. From that time forward, the medical profession came to believe in the germ theory of disease and in vaccinations to prevent disease. Within two years, anthrax was virtually eliminated from the French cattle and sheep industry.

Flores aimed the moral of his story at entrepreneurs: if you want to make an innovation that people will care about and value, look for the dead cows.

Dead Cows in Markovian Queueing Networks

Network-of-queues models were very attractive for computing systems and networks. They directly represent systems of servers in which jobs circulate. When a job arrives at a server, it waits in queue and then receives an interval of service before departing for another server. The state of the system at any given time is a vector $\boldsymbol{n} = (n_1,...,n_K)$ saying how many jobs are queued at each server. Randomness arises in these systems because the length of service during a visit to a server, and the next server visited after a departure, are not known. The randomness is described by a service distribution at each server and by a probability distribution of next server after a departure.

The traditional stochastic model (SM) for such systems assumes that the process by which the system moves through its states is Markovian: the successive service times are independent, successive transitions between servers are independent, the service distributions are exponential, and the system reaches a steady state.

A few tests during the 1960s and 1970s of the fit between these models and the throughput and response times of real systems were highly encouraging. For example, in 1965 Alan Scherr of MIT showed that the simple 2-server "machine repairman" model could be used to predict quite accurately the throughput and response time of the first time-sharing systems. Forest Baskett observed similar fits between models and measurements on systems at the University of Texas in the early 1970s.

Although they had the right structure, queueing network models were unattractive because of the high computation needed to calculate performance quantities with them. In the 1950s, Jackson, and again in the 1960s Jackson, Gordon, and Newell showed that the equilibrium state probability distribution of a network of queues model decomposed into a product form when arrivals and service times were all exponentials. Although the product form was much simpler to evaluate than numerically solving the balance equations, the computations were still intractable for all but the smallest systems. Thus testing the model was limited to small systems. In 1971, Jeff Buzen demonstrated a breakthrough: a quadratic algorithm for evaluating the product form. Suddenly performance analysts could compare models with large, real systems. In a long line of experimental studies, performance analysts concluded that these models would typically get throughput correct to within 10% of the observed value, and response time to within 25% of the observed value. Soon thereafter blossomed the now-flourishing industry of system performance evaluation and capacity planning.

But despite their empirical success, these models presented a serious problem. While performance analysts repeatedly found users interested in their queueing models, they constantly faced skepticism because no one trusted the models' assumptions. The models assumed that a system was in equilibrium; measurements in real systems showed constantly shifting measures and parameters at different times of day and days of week. The models assumed that inter-server job transitions were independent; in real systems transitions correlated with previous transitions. The models assumed that the service times at the servers were exponentially distributed; real systems had distinctly non-exponential service distributions, including many with very long tails. This presented a troubling paradox: the real world of computing systems consistently violated all the model assumptions, and yet the models agreed remarkably closely with observed throughput and response time.

This paradox was not a simple annoyance; it was standing in the way of business. Jeff Buzen and his partners, who were establishing a company (BGS Systems) to build and market performance prediction and capacity planning tools for the computing industry, knew this first hand. Distributed computing systems and networked systems were increasingly common and the performance and capacity questions for these systems were high on the minds of designers and users. Business executives were prepared to invest significantly in performance prediction and capacity planning -- their customers demanded it -- and yet they felt it unsafe to invest in technology based on what appeared to be black magic, technology whose limits were not understood.

To sidestep the skepticism, pragmatists pressed an empirical argument: "So what if the model assumptions don't hold? We can show empirically that the equations from the models work well, and that should be good enough for practical performance analysts." But business people weren't buying that argument. Did the empirical models rest on still hidden, deep principles? Or where they ad hoc? No one knew what the limits of the models might be or for which real systems they might fail. In other words, dead cows littered the intellectual landscape of performance modeling.

The Birth of Operational Analysis

Several of us were concerned about this and had started independent searches for a "vaccine". Dick Muntz and John Wong published a paper in the 1974 Princeton Conference showing that some of the formulas, including those for utilization, response time, and throughput, held in the limit for very general networks [11]. Jeff Buzen and I started to discuss this in 1975. We were struck by the parallels in our thinking and decided to collaborate. Jeff had already drafted papers, published in 1976, about fundamental laws (again utilization, throughput, and response time) that were always true because of the way they were defined for collected data [2,3]. Jeff suggested the term "operational analysis" to differentiate the approach from stochastic analysis. With my students I worked on a series of technical reports applying this form of analysis to multiprogrammed virtual memory systems in 1975 [1,14,15,16]. Jeff and I published a series of papers taking operational analysis into queueing networks beginning in 1977 [4,5,6,7,13]. I followed up with versions for *American Scientist* in 1991 [8,9].

The operational approach goes to fundamentals. In validating models, analysts substitute measured, operational values of parameters for the model's stochastic parameters. We wondered whether the direct substitution of measured values might be equivalent to interpreting the product form as the solution to a broader class of systems than the Markovian assumptions suggested. Could we find another set of assumptions that give the same equations but apply to large classes of real systems?

Queueing theory gives various limit theorems relating basic quantities for systems in equilibrium. For example, the utilization of a server is the product of the arrival rate and the mean service time ($U = XS$). In those days, we would "prove" this by solving a system's equations for utilization as a function of time and then taking the limit as time becomes infinite. Because students found it difficult to follow the mathematical details of such a proof, we would also offer an "intuitive" explanation based on how we would verify the limit theorem through an experiment. The intuitive explanation was this. If we observe a server for a period of time T, we can measure the number of service completions, C, and the total busy time, B. We can then empirically define the utilization as $U = B/T$, the mean service time as $S = B/C$, and the throughput as $X = C/T$. But with these definitions, it is *always* true that $U = XS$. The limit theorem of stochastic theory becomes an operational law when applied directly to the measured data.

We quickly found that several other limit theorems are also laws in the same way. For example, Little's law is $N=RX$, for mean number N in the system and mean response time R of the system. The forced flow law in a network is $X_i=XV_i$ where X_i is the throughput at server i, X the system throughput, and V_i the mean number of visits by a job to server i. The time sharing response time law is $R=N/X\text{-}Z$, where N is the number of users and Z is the average think time between submitting new commands. The memory space-time law says $XY=M$, where Y is the mean space-time per job and M is the total amount of memory used by jobs.

Jeff and I decided to see if we could start there, rather than finish there. Instead of concluding that $U=XS$ is true of systems in the limit, why not start with the observation that $U=XS$ is *always* true because of the way we define our measurements? In other words, $U=XS$ is a law that holds for all observation periods, including but not limited to those in Markovian equilibrium.

We found this so compelling that we then asked: Can we build a queueing theory that starts from the operational laws and avoids making any Markovian or equilibrium assumptions? To our delight we were able to do this. Operational Analysis (OA) became a vaccine whose new interpretation of systems prevented the death of trust in models (the cows).

The Fundamental Assumptions of Operational Analysis

We insisted that all assumptions we would make in our theory be operational: meaning that one can design an experiment to observe all the quantities we define. The experimental measurements would always be calculated in a given, arbitrary observation period of length T.

We insisted on testability, not because we advocated that everything be tested, but because we wanted the theory to be founded on behaviors that people can easily visualize. We often made an analogy with an old argument in physics where field theory replaced action-at-a-distance. One could explain the electrical attraction of two charged particles as an action over a distance (Coulomb's law); or one could say that one particle moves in an electric field of the other. The electric field was operational: one can imagine a small particle placed at any point in the field, and from the field direction and intensity one could say how the particle would move (a Δx in the next Δt). We fully realized that the large state spaces of networked systems would preclude actually testing all the assumptions, but we wanted to state them in a way that anyone wishing to understand the experiment would know exactly what we meant.

For queueing systems, the states $\boldsymbol{n}$=(n_1,...,n_K) are vectors giving the number of jobs queued up at each server. Stochastic modeling assigns an equilibrium probability $p(\boldsymbol{n})$ to each state. Operational analysis instead interprets the $p(\boldsymbol{n})$ as the proportions of time that the system spends in state $\boldsymbol{n}$. We called these $p(\boldsymbol{n})$ the *state occupancies*.

We re-formulated the familiar balance equations among the equilibrium $p(\boldsymbol{n})$ into balances of state transitions: entries=exits. We called this assumption *flow balance*. Entries and exits from states are observable and can be measured. Because the numbers of entries and exits need not match, we said that it is only an assumption that they are equal. We calculated the error that would arise in a solution of the balance equations when flow is not balanced in the real system. We showed that in long observation periods of systems with finite state spaces, the error caused by flow balance is negligible.

Because we wanted the balance equations to conform to the operational laws, we wanted the state transitions of the system to coincide with job completions at the individual servers. In other words, we wanted state transitions to correspond one-one with inter-server job transitions. This was easily accomplished with a second assumption: that each state change is caused by exactly one job completion. We called this assumption *one-step behavior*. As with flow balance, we calculated the error caused by this assumption and showed that in most real systems the error would be negligible.

With these two assumptions, the balance equations are mathematically identical to the equilibrium state probability equations of the same system interpreted as Markovian. We needed a third operational assumption to reduce these equations to the same form from which Jackson, Gordon, and Newell obtained the product form solution. This happened when we assumed that the rate of transitions between two states is identical to the rate of job-flow between the servers causing the transitions. We called this assumption *homogeneity*. As with the other two assumptions, we could calculate the error caused by this assumption. Unlike the other two assumptions, however, we could not show that the homogeneity error is negligible. In fact, in some systems, homogeneity introduces considerable error.

Under the homogeneity assumption, the configuration of queue lengths in the rest of the system does not affect the completion rate of a server, and hence that rate can be measured by studying the server in isolation from all other servers. Thus the homogeneity assumption is equivalent to an assumption that a server has the same completion rate (for a given queue length) in an on-line measurement as it will in an off-line measurement. For this reason we also called the homogeneity assumption the "on-line equals off-line" assumption.

Taken together, the three assumptions allowed us to reduce the balance equations and conclude that the state occupancies obey the same product form structure as had been found by Jackson, Gordon, and Newell.

We thus arrived at the same mathematical form as the Markovian theory, but with a completely operational interpretation. In addition to dealing directly with measured parameters, the operational interpretation can be applied in any finite observation period. It allows us to calculate the error caused by any of the three key assumptions. We had found a way to formulate queueing network theory so that the product form solution holds for finite intervals in which the system is flow balanced, one step, and homogeneous. Therefore, all the algorithms for solving product form networks could be used with confidence in many practical situations where their validity was dubious according to Markovian assumptions.

We found that operational analysis was less satisfactory for more complex systems such as an M/G/1 queue. Although we were able to formulate operational assumptions for the famous Pollaczek-Khinchtine formula for the mean queue length, the assumptions were more complex hard to understand. The mathematics of transform analysis, which are well known in SM, got to the same result more quickly.

Controversy

The new theory attracted a lot of attention -- from strong praise to strong criticism. In 1979, Ken Sevcik summarized the best and worst as follows [12]:

> "OA offers nothing but tautologies."
>
> "OA makes SM obsolete."
>
> "OA is a smokescreen for trivially deriving the obvious from the known."
>
> "SM is a security blanket used to smother intuition by those who lack it."

The most popular criticism focused on the homogeneity assumption, which the critics believed to be fundamentally equivalent to the exponential assumption. Ken attacked this criticism head on. In 1979 (with Maria Klawe) he gave several examples of deterministic systems that are flow-balanced, one-step, and homogeneous -- but obviously not Markovian. That was a turning point in the acceptance of operational analysis as a valid alternative to the traditional queueing theory. Many skeptics came on board after that. Ken drew several conclusions about the debate in 1979:

> OA invokes a different level of abstraction from the SM: the two systems have the same symbols but interpret them differently. SM refers to probabilistic ensembles of system behaviors; OA refers to one behavior at a time. OA is more obviously relevant to real systems than SM. OA generates confidence in applying models by offering assumptions that are understandable and testable. OA and SM are complementary approaches. OA offers much that is new; SM isn't obsolete.

After that, the debate became more philosophical. What does it mean to model? How are the approaches of SM and OA to creating and interpreting system models the same? Different? A hot example of this kind was the use of the models for performance prediction. The traditional practice of SM was to assume that the same stochastic process governs both the base and future observation periods. Therefore, one estimates the parameters from data in the base period and extrapolates them to the future period. Jeff and I argued that this practice didn't depend on an underlying SM. All it depended on was extrapolation of parameters to the future period. In any observation period, the computation of performance metrics from parameters uses algorithms that do not care whether the product form solution comes of operational or stochastic assumptions.

In 1981 I was present at a debate between Jeff Buzen and his critics. Neither the OA believers nor the SM believers were able to muster any argument that would change minds. Afterwards I wrote a fable to satirize the debate and suggest that the two sides may never come to an accord. A copy is attached as an appendix.

Despite their differences, the OA and SM believers did have one major point of agreement: most everyone found it *much* easier to teach queueing networks to beginning students when starting with the operational interpretation. I was able to teach queueing network basics to undergraduate computer science students in about two weeks of an operating systems class, compared to almost a whole course for Markovian theory. By thinking in the operational framework, my OS students developed a much better "feel" for how the models worked and their scopes of applicability. Ken Sevcik experienced the same thing with his students. Jeff Buzen experienced it in teaching his clients how the models work and why they can be trusted. Operational analysis gave an indisputable edge to teaching, understanding, and communicating about queueing networks. Because of this Ken embraced operational analysis to explain queueing theory in his best-selling book, which became the leading book in the field for many years [10]. More recent authors, such as Menascé and Almeida, have adopted operational analysis as their pedagogic tool for the same reason.

Salute

So I salute Ken Sevcik, whose insight at a critical point turned the tide in our favor and showed the skeptics that homogeneity was indeed a new assumption, more primitive and broader in scope than Markovian assumptions. Ken helped clear the field of its dead cows.

An Historical Footnote

When we formulated operational analysis, "queueing" had two "e's" in it. Microsoft Office spell checker now claims that "queuing" is the proper spelling. I tell recalcitrant editors that queueing is the only word in English with five consecutive vowels. So far this argument has prevailed.

References (Published)

1. Balbo, Gianfranco, and Peter Denning. "Homogeneous approximations of general queueing networks", *Proc. Int'l Symposium on Computer Performance Measurement, Modelling, and Evaluation*, North-Holland Publishing Co. (1979), in Vienna, Austria.

2. Buzen, Jeffrey. "Fundamental laws of computer system performance," *Proc IFIP-SIGMETRICS International Symposium on Computer Performance modeling, Measuremetn, and Evalaution*, Cambridge, MA (March 1976), 200-210.

3. Buzen, Jeffrey. "Operational Analysis: The key to the new generation of performance prediction tools," *Proc. IEEE COMPCON 76*, Washington, DC (September 1976), 166-171.

4. Buzen, Jeffrey, and Peter Denning. "Measuring and calculating queue length distributions", *IEEE Computer 13*, 4 (April 1980), 33-44.

5. Buzen, Jeffrey, and Peter Denning. "Operational treatment of queue distributions and mean value analysis", *Computer Performance 1*, 1 (June 1980), 6-15.

6. Denning, Peter, and Jeffrey Buzen. "Operational analysis of queueing networks", *Proc. 3rd Int'l Symposium on Modelling and Performance Evaluation of Computer Systems*, North-Holland Publishing Co. (1977), 151-172.

7. Denning, Peter and Jeffrey Buzen. "The operational analysis of queueing network models", *Computing Surveys 10*, 3 (September 1978), 225-261. Reprinted in *CMG Transactions*, Summer 1994, 29-60.

8. Denning, Peter. "Queueing in networks of computers." *American Scientist 79*, 3 (May-June 1991), 206-209.

9. Denning, Peter. "In the Queue: Mean Values." *American Scientist 79*, 5 (September-October 1991), 402-403.

10. Lazowska, Edward, John Zahorjan, G Scott Graham, and Kenneth Sevcik. *Quantitative System Performance*. Prentice-Hall (1984).

11. Muntz, Richard, and John Wong. "Asymptotic properties of closed queueing network models. *Proc. 8th Princeton Conference on Information Sciences and Systems*, Dept EECS, Princeton University (1974), 348-352.

12. Sevcik, Kenneth C, and Maria Klawe. "Operational Analysis versus Stochastic Modelling of Computer Systems." *Proc. of Computer Science Statistics: 12th Annual Symposium on the Interface*, Waterloo, Canada (May 1979), 177-184.

References (Unpublished Technical Reports)

13. Buzen, Jeffrey, and Peter Denning. "Operational analysis of Markov Chains." File memorandum PD78.1 (May 1978).

14. Denning, Peter, and Kevin Kahn. “Some distribution free properties of throughput and response time”, Purdue University, CS Dept., CSD-TR-159 (May 1975), 28pp.

15. Denning, Peter. “Operational laws of system performance,” File Memorandum PD75.6 (June 1975), 8pp.

16. Denning, Peter. “Asymptotic properties of multiprogrammed response,” File Memorandum PD75.11 (August 1975), 13pp.

APPENDIX --OPERATIONAL ANALYSIS: A FABLE

Peter J. Denning

30 Jan 1991

Operational queueing theory was controversial among queueing theorists. A popular criticism was that the operational assumption of homogeneity --- service rates of servers do not depend on total system state -- was nothing more than an exponential service-time assumption in disguise. That criticism was neatly dispelled by Ken Sevick and Maria Klawe, whose examples of operationally-deterministic systems in no sense satisfied an exponential service time assumption, but satisfied product form solutions. Another criticism was that one cannot make predictions of a future system's performance without assuming the present and future systems are manifestations of the same underlying stochastic process. Buzen said that stochastic processes had nothing to do with it; he argued that prediction in practice operates as a process of extrapolating present to future parameter values and then using a validated model to calculate future performance measures. Such logic did little to assuage some critics, who maintained that operational analysis denied the existence of stochastic processes.

In 1981, I witnessed a debate between Buzen and his critics. I was struck by the symmetry of their arguments. Each started with his domain as the ground and claimed that the other was in effect performing unneeded, error-inducing mappings to get to the same answer. They were both describing the same loop from different angles! This prompted me to write the following little fable.

A Tale of Two Islands

Once upon a time there were two islands. The citizens of Stochasia had organized their society around a revered system of mathematics for random processes. The citizens of Operatia had organized their society around a revered system for experimentation with nondeterminate physical processes. Both societies were closed. Neither would ever have known of the other's existence, had it not been for the events I shall now describe.

At a moment now lost in the mists of antiquity, a great sage of Stochasia posed this problem: Given a matrix of transition probabilities, find the corresponding equilibrium probability distribution of occupying the possible states. He worked out the solution, which he engraved on stones. Ever since, whenever they encounter a problem in life, the Stochasians phrase it in these terms and, using the stones, they find and implement its solution.

At a moment now lost in the mists of antiquity, a great sage of Operatia posed *this problem: Having observed a matrix of transition frequencies, calculate the corresponding distribution of proportions of time of occupying the possible states. He worked out the solution, which he engraved on stones. Ever since, whenever they encounter a problem in life, the Operatians phrase it in these terms and, using the stones, they find and implement its solution.*

In a recent time there was an anthropologist who specialized in islands. He discovered these two islands from photographs taken by an orbiting satellite. He went to visit Stochasia, where he learned the secrets of their stones. He also visited Operatia, where he learned the secrets of their stones.

Struck by the similarities, the anthropologist asked the elders of each island to evaluate the approach used by the other island. In due course, each island's elders reached a decision.

The elders of Operatia told the anthropologist: "The Stochasians are hopelessly confused. They have developed a highly indirect approach to solving the problem posed by our great sage. First, they transform the problem into an untestable domain by a process we would call 'abstraction'. Using their stones, they find the abstract answer corresponding to the abstract problem. Finally, they equate the abstract answer with the real world by a process we would call 'interpretation'. They make the audacious claim that their result is useful, even though the two key steps, abstraction and interpretation, can nowise be tested for accuracy. Indeed, these two steps cannot be tested even in principle! Our stones tell us elegantly how to calculate the real result directly from the real data. No extra steps are needed, and nothing untestable is ever used."

The elders of Stochasia told the anthropologist: "The Operatians are hopelessly confused. They have developed a highly indirect approach to solving the problem posed by our great sage. First, they restrict the problem to a single case by a process we would call 'estimation'. Using their stones, they estimate the answer corresponding to their estimate of the problem. Finally, they equate the estimated answer with the real world by a process we would call 'induction'. They make the audacious claim that their result is useful, even though the two key steps, estimation and induction, are nowise error free. Indeed, these two steps cannot be accurate even in principle! Our stones tell us elegantly how to calculate the general answer directly from the parameters. No extra steps are needed, and nothing inaccurate is ever used."

The anthropologist believed both these arguments and was confused. So he went away and searched for new islands.

Some years later, the anthropologist discovered a third island called Determia. Its citizens believe randomness is an illusion. They are certain that all things can be completely explained if all the facts are known. On studying the stones of Stochasia and Operatia, the elders of Determia told the anthropologist: "The Stochasians and Operatians are both hopelessly confused. Neither's approach is valid. All you have to do is look at the real world and you can see for yourself whether or not each state is occupied. There is nothing uncertain about it: each *state is or is not occupied at any given time. It is completely determined."*

Later, he told this to an Stochasian, who laughed: "That's nonsense. It is well known that deterministic behavior occurs with probability zero. Therefore, it is of no importance. How did you find their island at all?" Still later, he told this to an Operatian, who laughed: "I don't know how to respond. We have not observed such behavior. Therefore it is of no importance. How did you find their island at all?"

The anthropologist believed all these arguments and was profoundly confused. So he went away and searched for more new islands. I don't know what became of him, but I heard he discovered Noman. (Noman is an island.)