# The Power Function as a Performance and Comparison Measure for ATM Switches *

Christos Kolias and Leonard Kleinrock

Department of Computer Science
University of California at Los Angeles
Los Angeles, CA 90095-1596
U.S.A.

## Abstract

In this paper we apply the notion of *Power*, as has been defined by one of the authors for a general communications system, to ATM switching systems. In general, the Power of a system synthesizes basic performance metrics such as the system's throughput, mean delay and packet loss. This integrated measure can be used to characterize the performance of a system (e.g., identify its optimal operating point) or even to compare it to other systems. We utilize the power function to systematically quantify and compare the performance of ATM switches.

## 1 Introduction

The performance of a communications system is very critical and fundamental in designing, evaluating or comparing it to other systems. We consider a communications system where packets (or messages) arrive, receive service and depart. Let $\lambda$ be the *input load* applied to the system and $\bar{x}$ denote the mean service time (received by a packet). The following fundamental system performance characteristics are defined:

- *throughput*, $\gamma = \gamma(\lambda)$, the rate at which packets receive service (i.e., depart from the system),
- *mean delay* (response time), $T = T(\lambda)$, the average time (queueing and service) spent by a packet while it resides in the system,
- *blocking probability*, $P_B$, the probability that an arriving packet gets rejected (denied access) by the system. $P_B$ is a function of $\lambda$ and can be a function of other parameters too, depending on the system description, such as buffer size, number of servers, bandwidth.

From these principal parameters a number of additional descriptors may be obtained such as the *utilization* factor, denoted by $\rho$ $(0 \leq \rho \leq 1)$, which indicates the proportion of time the system is busy (i.e., serving packets).[1] Furthermore, using Little's result, the average queue length (number of packets queued) $L$ or the average number in system $\overline{N}$ can be obtained (i.e., $\overline{N} = \gamma T$).

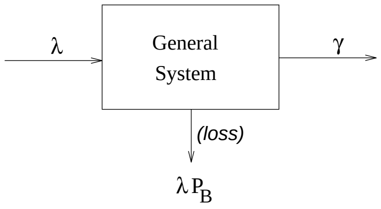


Figure 1: A schematic description of the performance characteristics of a general system.

The original set of performance measures constitute a quantitative basis and a set of common criteria for describing a system's behavior and evaluating its performance. Fig. 1 depicts a general system and some of its basic performance metrics. From the same figure note that $\lambda P_B = \lambda - \gamma$. It is often the case that a system cannot reach its highest capacity thus not yielding a 100% throughput. That means that the system reaches its maximum throughput at input loads where $\lambda \bar{x} < 1$. Exceeding that critical load drives the system to overloaded situations (i.e., delay and queue length grow to infinity). Below we show an example of a throughput-input load function

$$\gamma = \begin{cases} \lambda, & \lambda \leq \gamma_{max} \\ \gamma_{max}, & \lambda \geq \gamma_{max} \end{cases} \quad (1)$$

where $\gamma_{max}$ represents the maximum input load that can be handled (and hence the maximum throughput achieved) by the system without causing any instability. Thus, in order for a system to be stable we require that $\lambda < \gamma_{max}$. The above example, shown graphically in Fig. 2, refers to the ideal case (for a no-loss system, i.e., $P_B = 0$); normally throughput degrades for $\lambda > \gamma_{max}$, unless some type of flow control is implemented. As $\lambda$ (or $\gamma$) approaches $\gamma_{max}$ the system becomes saturated. It is

*This research was supported by the DARPA/ITO under contract DABT63-93-C-0055, *The Distributed Supercomputer Supernet - A Multi-Service Optical Intelligent Network.* Web URL *http://millennium.cs.ucla.edu/~ck*

[1] For a queueing system with $m(< \infty)$ servers, e.g., $G/G/m$, it holds that $\rho = \gamma\bar{x}/m$ and it is required $\rho < 1$ in order for the system to be stable (ergodic).

then apparent that $\gamma_{max}$ is the critical value and finding it is precisely the key point in a throughput analysis.

Fig. 3 shows a mean delay profile. The minimum mean delay measured in a system is that of the mean service time $\bar{x}$. For the above example of throughput, $T$ is expressed as

$$T = \bar{x} + \frac{A}{\gamma_{max} - \gamma}$$

Thus, as $\gamma \rightarrow \gamma_{max}$ then $T \rightarrow, \infty$ where we assume that $A$ varies slowly with $\gamma$.

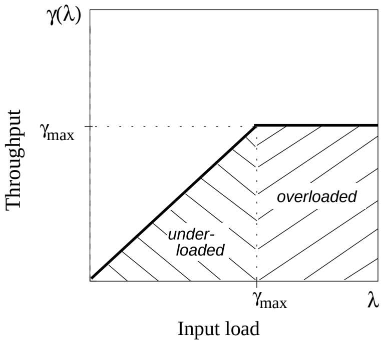


Figure 2: The (ideal) throughput characteristic of a system.

How these characteristics, namely throughput, mean delay and blocking, interplay is the focus of our discussion. Recognizing the practical importance of evaluating ATM switching systems our objective is to collectively use these main performance measures of our interest in order to study the performance across various ATM switches.

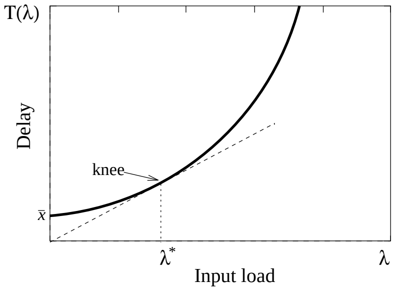


Figure 3: The delay characteristic of a system.

## 2 The simple Power Function

The fundamental need to characterize the overall performance of a communications system led one of the authors [7], [8] to define and investigate a homogeneous performance measure known as *Power* [2] [1]. The power of a system synthesizes the system's various performance metrics and characteristics such as: throughput, mean delay, blocking probability. It can be extended to even include other factors such as delay variance, cost) as to make it more robust and accurate. As a function, power can be used to identify those operating point where a system delivers its best performance. Power combines "competing" performance measures such as throughput and delay or delay and blocking. As we increase the applied input load, throughput increases (see Eq. (1)) but delays become longer (as the system tends to saturate). There is clearly a tradeoff between those two measures. The notion of power proves to be useful in addressing this tradeoff issue. It appears as a natural measure for characterizing real-time traffic such as video and speech, whose efficient transmission requires, simultaneously, high throughput and low delays.

For simplicity, in what follows we will use throughput to measure the system's utilization (which is $\gamma\bar{x}$) and we will assume that $T$ is normalized over $\overline{x}$, (i.e., $T/\bar{x}$).[3] For consistency, we also make $\lambda$ unitless, with $0 \leq \lambda \leq 1$. The Power of a no-loss system ($P_B = 0$) is then defined as follows [7]:

**Definition 1** *The simple Power function, $P \triangleq P(\lambda)$, of a (no-loss) system is defined as the ratio of the system's throughput, $\gamma$, over the system's mean (normalized) delay, $T$:*

$$P = \frac{\gamma}{T} \tag{2}$$

Note that $0 \leq P \leq 1$, since $\gamma \leq 1$ and $T \geq 1$. In fact, $P = 1$ for the $D/D/1$ and $M/M/\infty$ queueing systems.

Let us consider any $T(\lambda)$ that is continuous and nondecreasing. Then $P$ is maximized for that value of $\lambda$, which we denote by $\lambda^*$, where $P^{(1)}(\lambda) = \frac{dP(\lambda)}{d\lambda} = 0$ and $\frac{d^2P(\lambda)}{d\lambda} < 0$. As a result it holds [7] that $P$ is maximized when $\frac{T^{(1)}(\lambda)}{\gamma^{(1)}(\lambda)} = \frac{T(\lambda)}{\gamma(\lambda)}$ According to Eq. (1) this last equation can be written as

$$\frac{dT}{d\lambda} = \frac{T}{\lambda} \tag{3}$$

As Fig. 3 shows, the load at which power is maximized, namely $\lambda^*$, can be found by taking the tangent to $T(\lambda)$ from the origin,[4] in which case: $\frac{dT}{d\lambda} |_{\lambda=\lambda^*}= \frac{T(\lambda^*)}{\lambda^*}$.

We denote by $P^*$ the maximum power achieved by the system, i.e., $P^* = P(\lambda^*)$. Note also that for $\lambda \geq \gamma_{max}$, $P = 0$. That follows from the simple fact that for any $\lambda \geq \gamma_{max}$ the mean delay $T$ becomes infinite. Therefore we consider only cases where $\lambda < \gamma_{max}$, i.e., non-saturation or underloaded situations (Fig. 2). By definition, $T^* \triangleq T(\lambda^*)$ and $\gamma^* \triangleq \gamma(\lambda^*)$. We further have that the optimal mean number in system is $\overline{N}^* = \gamma^* T^* = (\gamma^*)^2/P^*$.

From our discussion above it can be further argued that the larger the power, the higher the "performance" of the system is.

## 3 The Power Function in ATM Switching

As we will later see, the power function has a great applicability and significance in the area of ATM switch-

[2] The notion of power for a communications system takes its analogy [1] from physics, where throughput corresponds to energy and delay to time.

[3] As we will later see this is a quite reasonable assumption as $\bar{x} = 1$ for ATM switches.

[4] This tangency point is referred to as the "knee" of the curve.

ing since the performance of an ATM switch is described precisely by the same set of performance measures as a general system.

In general, an ATM switches is modeled as a timeslotted queueing system where a timeslot is the time it takes for a cell to be switched, and thus it represents the service time received by a cell, i.e., $\bar{x} = 1$. The switch models we study are $N \times N$ switches ($N$ inputs, $N$ outputs), where $N$ is very large (theoretically $N \rightarrow \infty$). We assume that cell arrivals at the inputs of the switch are governed by a Bernoulli distribution with parameter $\lambda$. Furthermore, we assume that incoming traffic is uniformly distributed among the output ports. Also in this section we consider switches with infinite size queues, i.e., $P_B = 0$. The throughput of a generic ATM switch is described exactly by Eq. (1).[5] Delay is measured as the total (queueing and switching) time spent by a cell while it is in transit inside the switch. The simple power function is then expressed (for underloaded situations[6]) as:

$$P = \frac{\lambda}{T} \tag{4}$$

We restrict our attention to three architecturally different, but with common characteristics, switching systems.

## 3.1 Crossbar switches with Multiple Input-Queueing

We first examine the power function for switches that use Multiple Input- Queueing [3]. Every input port has $m(\leq N)$ independent queues (Fig. 4). Each of these queues stores incoming cells that are destined to a unique subset of output ports. This results in a $mN \times N$ crossbar switch[7] as depicted in Fig. 4 (each value of $m$ yields a different switching system). We allow all $mN$ queues to participate in the arbitration phase (and contend for output ports) at the beginning of each timeslot.

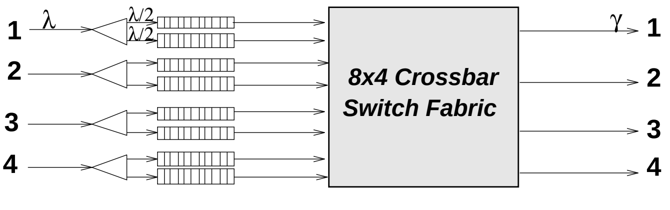


Figure 4: A $4 \times 4$ ATM switch with multiple input-queues ($m = 2$).

This scheme dramatically alleviates the *head-of-line* (HOL) problem [3]. In [4] we prove that the maximum throughput achieved by the switch is $\gamma_{max} = 1 + m - \sqrt{1 + m^2}$. In the same paper we also derive an exact expression for its mean delay $T$ as a function of $\lambda$ and $m$. In Fig. 5 we show the various mean delay curves for different values of $m$. Note that they start from $T = 1$ which is, again, the minimum delay, namely that of the service time, a cell experiences.

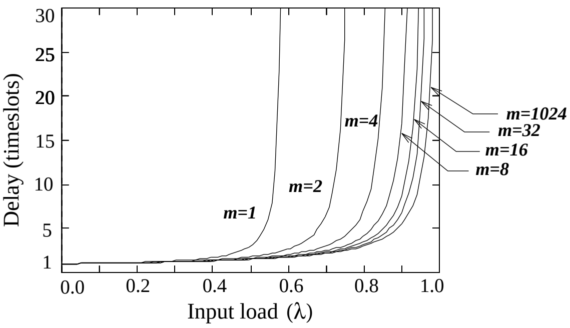


Figure 5: Mean delay *vs.* input load in the Multiple Input-Queueing scheme.

Since we have available our two main performance indices it behooves us to apply the power function, given by Eq. (4), to this queueing system, which gives

$$P = \frac{6\lambda(1-\lambda)}{11\lambda^4 - (32+6m)\lambda^3 + (48+30m)\lambda^2 - \dfrac{(\lambda^2 - 2(1+m)\lambda + 2m)}{-(24+48m)\lambda + 24m}}$$

In Fig. 6 we plot this power function for different values of $m$. As $m$ increases so does the power. We see from the same figure that the gain in power becomes diminishing as $m$ increases. We have numerically calculated the peak of each $m$-curve, which represents the maximum achievable power ($P^{(1)}(\lambda) = 0$) and tabulated the results in Table 1. From the same table and from Fig. 6 we see that in fact there is a little gain in performance going from $m = 128$ to $m = 1024$. This finding alone corroborates an earlier observation of ours [4] that we do not need as many queues per port ($m$) as the number of output ports, namely $N$ in order to achieve a very high performance.

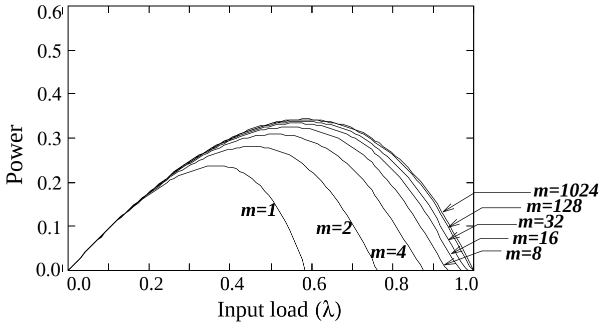


Figure 6: Power *vs.* input load in the Multiple Input-Queueing scheme.

If $\overline{N}$ represents the average number of cells that are present per input port (that is in all of the input's $m$ queues) then we have the following:

**Theorem 1** *For the Multiple Input-Queueing ATM switch, with m queues per input port, it holds that*

$$\lim_{m \rightarrow \infty} \overline{N}^* = 1 \tag{5}$$

[5] Note that by throughput we mean what is generally known as the efficiency of the system, that is, the utilization of the system or, in the case of switches, of the output ports.

[6] Note that $P \rightarrow 0$ as $\lambda \rightarrow \gamma_{max}$.

[7] An alternate design would be to consider $m$ ( $N \times N$) crossbar switching planes in parallel, where each queue would then feed a single plane.

| $m$ | Max. Power, $P(\lambda^*)$ | $\lambda^*$ |
|---|---|---|
| 1 | 0.2370 | 0.367 |
| 2 | 0.2809 | 0.454 |
| 4 | 0.3089 | 0.512 |
| 8 | 0.3251 | 0.546 |
| 16 | 0.3338 | 0.565 |
| 32 | 0.3384 | 0.575 |
| 128 | 0.3419 | 0.583 |
| 1024 | 0.3431 | 0.586 |

Table 1: Maximum Power in the Multiple Input-Queueing for different values of $m$.

*Proof.* From the equation expressing power we find

$$\lim_{m\to\infty} P = \frac{2\lambda(1-\lambda)}{2-\lambda} \qquad (6)$$

Taking $P^{(1)}(\lambda) = 0$ and solving for $\lambda$ we get $\lambda^* = 2 - \sqrt{2} \approx 0.586$, which gives $P^* = 6 - 4\sqrt{2}$, as also shown in Table 1 (for $m = 1024$). It readily follows that $\overline{N}^* = (\lambda^*)^2/P^* = 1$ which is also the celebrated result for $M/G/1$ [7].[8] The physical interpretation is that the optimal mean number of cells per input port (in any of the $m$ queues) is exactly 1. □

In the next two subsections we consider $N \times N$ multiplane switches with $m$ crossbar fabrics that operate in parallel and which employ queues on both their inputs and outputs.[9]

### 3.2 Multiplane Switches with Single Input-Queues

Here we study the performance of a multiplane switch constructed from $m$ ($N \times N$) crossbar switches where an input queue is connected to all planes (Fig. 7). The HOL cell (the cell at the head position of its input queue) randomly chooses one of the $m$ switching planes to go through. Queues are required at the outputs to collect the cells routed through the various planes.

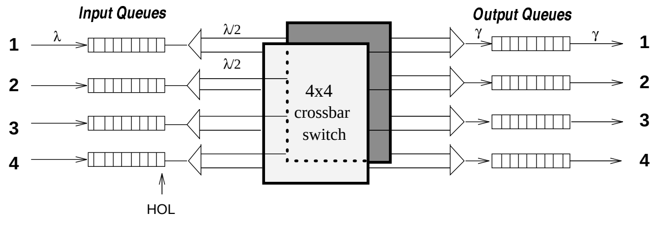


Figure 7: A $4 \times 4$ ATM switch with $m = 2$ switching planes ($4 \times 4$ crossbars) and Single Input-Queues.

In [6] we give analytical expressions for both the throughput and mean waiting time ($\overline{W}$) for this type of multiplane switch. Fig. 8 illustrates the mean delay $T = \overline{W} + 1$ as a function of the applied input load $\lambda$. Note that these mean delay curves start from $T = 2$ since this is the minimum delay a cell can experience, as it takes a cell one timeslot to get switched out from its input queue (and to its selected output queue) and another one to depart through its output port. The maximum throughput achieved is (cf. [6]) $\gamma_{max} = 1 + m - \sqrt{1 + m^2}$, a familiar already result.

Power is then expressed as follows:

$$P = \frac{6\lambda(1-\lambda)(m-\lambda)(2m-\lambda)}{-(14+\frac{3}{m})\lambda^5 + (47m + 44 + \frac{6}{m})\lambda^4 - m^2(72 + \frac{140}{m} +}$$

$$\frac{(\lambda^2 - 2(1+m)\lambda + 2m)}{+\frac{42}{m^2})\lambda^3 + 6m^3(6 + \frac{31}{m} + \frac{18}{m^2})\lambda^2 - 12m^3(7 + \frac{10}{m})\lambda + 48m^3}$$

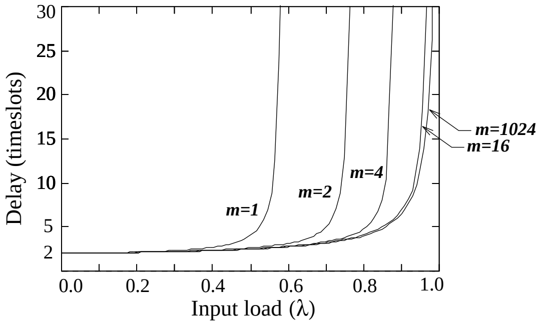


Figure 8: Mean delay *vs.* input load in the Single Input-Queues multiplane switch.

Fig. 9 exhibits the power vs. input load relationship for this type of multiplane switch. We notice that the curves for $m = 16$ and for $m = 1024$ are very close, leading us to the assessment that $m = 16$ is sufficient for implementing this multiplane switch. Table 2 which shows the maximum power for the various values of $m$ supports this conclusion.

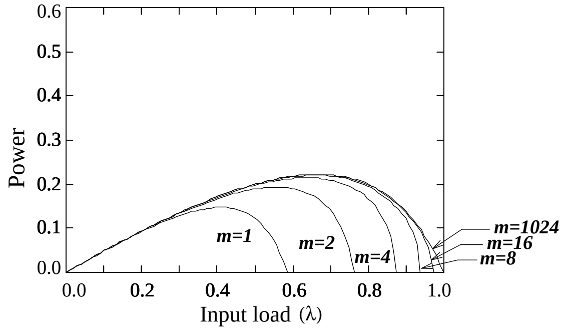


Figure 9: Power *vs.* input load in the Single Input-Queues multiplane switch.

Moreover, given that $\overline{N}$ is the average number of cells per input-output queue pair we have the following:

**Theorem 2** *For the multiplane ATM switch with Single Input-Queues and m switching planes, it holds that*

$$\lim_{m\to\infty} \overline{N}^* = 2 \qquad (7)$$

[8] It can be shown that $\overline{N^*} = 1$ is true also for a $Geom/G/1$ queue. We remind the reader that $Geom/G/1$ is the discrete analog of $M/G/1$.

[9] See [6] for a complete description of these multiplane switches.

| $m$ | Max. Power, $P(\lambda^*)$ | $\lambda^*$ |
|---|---|---|
| 1 | 0.1476 | 0.411 |
| 2 | 0.1931 | 0.557 |
| 4 | 0.2147 | 0.637 |
| 8 | 0.2205 | 0.660 |
| 16 | 0.2218 | 0.665 |
| 32 | 0.2221 | 0.667 |
| 128 | 0.2222 | 0.667 |
| 1024 | 0.2222 | 0.667 |

Table 2: Maximum Power in the Single Input-Queues multiplane switch for different values of $m$.

*Proof.* From the last equation expressing power we get

$$\lim_{m\to\infty} P = \frac{2\lambda(1-\lambda)}{4-3\lambda} \tag{8}$$

from which we easily find that $\lambda^* = 2/3$ yielding a maximum power of $P^* = P(\lambda^*) = 2/9$, as also evident from Table 2. We immediately have $\overline{N}^* = (\lambda^*)^2/P^* = 2$. □

The idea here is to "keep the switch busy" by having, on the average, one cell per (input and output) queue.[10]

## 3.3 Multiplane Switches with Multiple Input-Queues

Last we consider a multiplane switch where now each (of the $m$) planes is fed by a separate (and specific) queue from each input port. This yields a total of $mN$ input-queues (Fig. 10). In reality, it is equivalent of stacking up $m$ $N \times N$ identical and independent input-buffered ATM switches. Again, queues at the outputs serve the purpose of collecting the relayed cells.

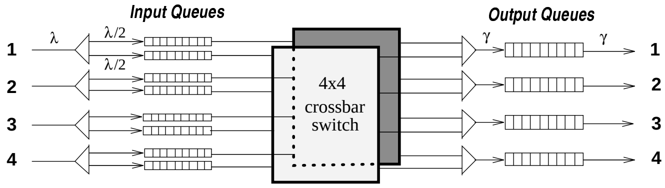


Figure 10: a $4 \times 4$ ATM switch with $m = 2$ switching planes ($4 \times 4$ crossbars) and Multiple Input-Queues.

Now, since the maximum throughput of each of these switches is known [2] to be $2-\sqrt{2}$ then the total maximum throughput for the multiplane switch is $\gamma_{max} = m(2-\sqrt{2})$. The obvious observation that follows is that $m = 2$ planes are enough to realize this type of switch since throughput cannot exceed 100%. In [6] we give an analytical expression for the mean delay, as a function of $\lambda$ and $m$, which we plot in Fig. 11.

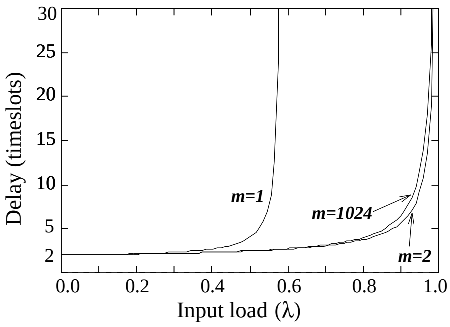


Figure 11: Mean delay *vs.* input load in the Multiple Input-Queues multiplane switch.

The following expression for the power of this multiplane switching system is derived:

$$P = \cfrac{6\lambda(1-\lambda)(m-\lambda)(2m-\lambda)}{-(14+\frac{3}{m})\lambda^5+(59m+38)\lambda^4-(126m+128)m\lambda^3+\cfrac{(\lambda^2-4m\lambda+2m^2)}{+(114m+216)m^2\lambda^2-(36m+168)m^3\lambda+48m^4}}$$

The power for this type of switch is of special interest. In Fig. 12 we plot the power function and we notice that power is not an increasing function of $m$ beyond $m = 3$ which yields the optimal power;[11] it gets slightly worse as $m$ grows. This rather erratic behavior is attributed to the fact that the mean delays tend to become larger with $m$ as the load shifts from the inputs to the outputs, which occurs as $m$ increases (more cells are switched to the outputs on a timeslot basis).

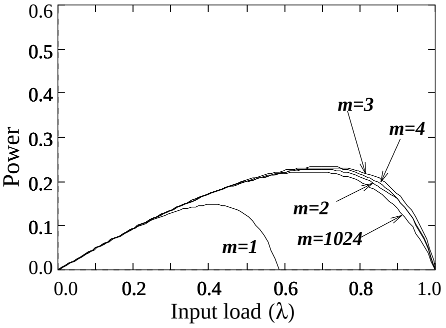


Figure 12: Power *vs.* input load in the Multiple Input-Queues multiplane switch.

Table 3 shows the peak values of power and their corresponding optimal input loads. This table complements our observations above. In fact we note that for $m > 7$ the maximum power is smaller than that for $m = 2$.

As in section 3.2 we have the following theorem:

**Theorem 3** *For the multiplane ATM switch with Multiple Input-Queues and $m$ switching planes, it holds that*

$$\lim_{m\to\infty} \overline{N}^* = 2 \tag{9}$$

*Proof.* Since

$$\lim_{m\to\infty} P = \frac{2\lambda(1-\lambda)}{4-3\lambda} \tag{10}$$

the proof is basically the same as in Theorem 2. Therefore $P^* = 2/9$ at $\lambda^* = 2/3$ (see also Table 3) and $\overline{N}^* = 2$. □

[10] This can be actually proven, if we consider the mean delays each of the input and output queues is contributing to the total delay $T$.

[11] $m$ need not be a power of 2, though it would be convenient for design purposes.

| $m$ | Max. Power, $P(\lambda^*)$ | $\lambda^*$ |
|---|---|---|
| 1 | 0.1476 | 0.411 |
| 2 | 0.2282 | 0.690 |
| 3 | 0.2347 | 0.703 |
| 4 | 0.2329 | 0.695 |
| 5 | 0.2312 | 0.689 |
| 6 | 0.2299 | 0.686 |
| 7 | 0.2289 | 0.683 |
| 8 | 0.2281 | 0.681 |
| 16 | 0.2253 | 0.674 |
| 128 | 0.2226 | 0.668 |
| 1024 | 0.2223 | 0.667 |

Table 3: Maximum Power in the Multiple Input-Queues multiplane switch for different values of $m$.

# 4 The Power function as a comparison measure

As we have indicated, another useful feature of the power function is that it is very attractive as a comparison measure. We use the power function not only to compare how a single switching architecture performs, e.g., for the various values of $m$, but also to compare the different switching architectures themselves.

Comparing Fig. 6 to Figs. 9 and 12 we see that multiple input-queueing actually achieves a better performance since it has a higher power. However, in the multiplane switches their peak power levels are obtained at higher input traffic loads $\lambda$, which may be preferable under heavy traffic conditions. Therefore, choosing a suitable design depends on the applied load at which we would like the switch to operate.

Comparing Figs. 9 and 12 we notice that they both offer comparative performance. However, since the first multiplane scheme is simpler in terms of hardware design and requirements (i.e., less hardware components) we might be inclined to choose that one. We also saw that the power $P$ is not always an increasing function of $m$ (see section 3.3).

# 5 The modified Power function

In this section we extend the concept of the power function as to include packet loss;[12] thus we consider ATM switches with finite size buffers. In an ATM switch, loss can occur when cells are forced to be dropped due to lack of buffer space. This loss contributes to the overall measured *cell loss ratio* which gives the fraction of the total number of cells transmitted that are lost or discarded (on a VCC basis). In section 1 we defined $P_B$ as the packet blocking probability. We use the same symbol to denote the cell drop probability inside a switch. It is clear that increasing the applied input load can lead to a significant increase of cell drop. It is then apparent that minimizing this cell blocking is an additional objective. The power function is then modified, in order to account for $P_B$, as follows [8]:

$$P_M = \frac{\gamma}{T}(1 - P_B) \tag{11}$$

where throughput is now expressed as

$$\gamma = \lambda(1 - P_B) \tag{12}$$

Note that the term $1 - P_B$ appears twice in the numerator of the modified power function; thus the emphasis on blocking is evident. This is a reasonable definition as we later discuss. Buffers operate as a dynamic flow control mechanism, they regulate the incoming traffic to the switch. Although dropping cells is definitely an undesirable action it is unavoidable given the finite capacity of the buffers.[13]

A more comprehensive and complete study of ATM switches with finite multiple input queueis considered in [4]. We use this example to precisely demonstrate the applicability of the modified power function, as given in Eq. (11), in ATM switching systems. We assume that the total buffer capacity per input port is $b$, thus a switch with $m$ queues per port mean that each queue can hold up to $K = b/m$ cells.

Let then $p_K$ be the probability that there are $K = b/m$ cells in the queue (i.e., a full queue); then $P_B = p_K$. The queue length distribution [9] is then given by

$$p_k = \frac{(1 - P_B)p_k^\infty}{\sum_{j=0}^{K-1} p_j^\infty}, \qquad k = 0, 1, ...K - 1 \tag{13}$$

where $p_k^\infty, k \geq 0$ is the queue length distribution in the multiple input-queueing switch with infinite buffers and is given in [4]. Then

$$P_B = \frac{(\lambda^2 - 2(1+m)\lambda + 2m)\left(\frac{\lambda^2}{2(1-\lambda)(m-\lambda)}\right)^K}{2(1-\lambda)(m-\lambda) - \lambda^2\left(\frac{\lambda^2}{2(1-\lambda)(m-\lambda)}\right)^K} \tag{14}$$

The mean queue length is: $L = \sum_{k=1}^{K} kp_k$. Using Little's result we obtain the mean delay for the finite capacity multiple input-queueing switch, which is

$$T = \frac{mL}{\gamma} + 1 \tag{15}$$

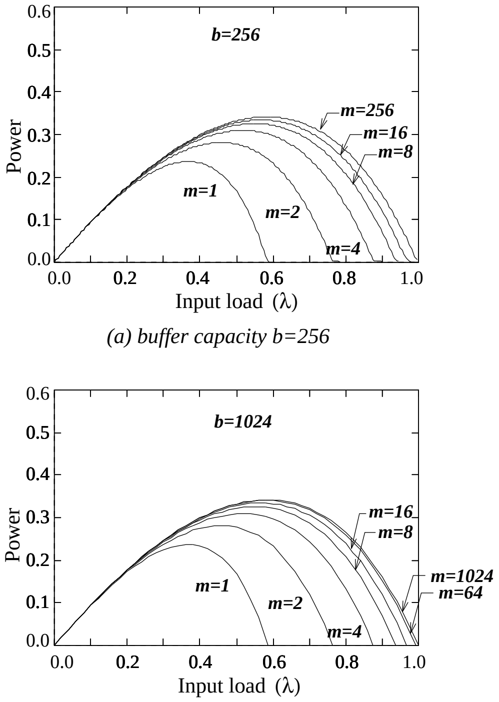
From Eqs. (11)-(15) we obtain the power for this finite-buffered switch. Figs. 13 (a)-(b) display this modified power function for two buffer capacities, $b = 256$ and 1024. Note again that a buffer capacity of $b$ means an individual queue size of $K = b/m$. Thus we are allowed to plot those curves only for $m \leq b$.

These curves support our earlier observation that blocking, namely the cell drop probability $P_B$ should play an important role in estimating the power in Eq. (11).

[12] Generally, packet loss in a system can be caused due to insufficient buffer capacity (e.g., $M/M/1/K$) or due to some (free) server unavailability (e.g., $M/M/m$) or both (e.g., $M/M/m/K$).

[13] Generally, a cell loss within the range $10^{-8}$ – $10^{-12}$ is acceptable for in ATM networks.

The power for the $m = b$ switching system increases too, thus it presents no unusual behavior. As a side comment we mention that in Fig. 13(a) the curves corresponding to $m = 64$ and $m = 256$ overlap. In Fig. 13(b) we notice that moving up from $m = 64$ to 1024 has little effect on improving the achieved power (consistent with the infinite-buffer case).



*(a) buffer capacity b=256*

*(b) buffer capacity b=1024*

Figure 13: Modified Power, $P_M$, *vs.* input load in the Multiple Input- Queueing scheme.

The interested reader may compare Fig. 13(b) to Fig. 6 and conclude that a buffer capacity of $b = 1024$ is then appropriate to implement a finite-buffered multiple input-queueing ATM switch and still achieve the (theoretical) performance offered by that switch with infinite buffers [4]. We can further state that even $b = 256$ yields a very good approximation of the infinite buffer case without any significant performance degradation.

# 6 Conclusion

We have applied the notion of *power* to ATM switching systems. We have investigated this power function under a variety of similar switching architectures that deal with the head-of-line problem and offer considerable performance enhancements. For the switching examples we considered, analytical expressions were given for various metrics involved. The same type of evaluation and comparison based on the power function can be extended to any switching systems. If no analytical results are available, one may use actual measurement or simulation statistics in order to find the power of a switching system. We studied the power function for systems with and without cell loss.

We recognize that another important performance measure which can be also critical is the variance of the cell delay. For instance we can add the term $C_s^2 + 1$ in the denominator of Eq. (11), where $C_s^2$ is the coefficient of variance for the delay. We can use the same approximation model as for the queue length distribution, namely the $Geom/Geom/1$ queue in order to obtain the delay distribution and its second moment.

The beauty of the power function derives from its simplicity and ease of implementation and understanding, while powerful in its ability to capture the overall performance of a system. We have also demonstrated that power is extremely useful, not only as a leading performance index, but also as a comparison tool.

# References

[1] A. Giesler, J. Hänle. A. König and E. Pade, "Free Buffer Allocation- An Investigation by Simulation", *Computer Networks*, **2**(3), pp. 191-208, July 1978.

[2] M. J. Karol, M. G. Hluchyj and S. P. Morgan, "Input versus Output Queueing on a Space-Division Packet Switch", *IEEE Trans. Commun.*, **COM-35**(12), pp. 1347-56, Dec. 1987.

[3] C. Kolias and L. Kleinrock, " Throughput Analysis of Multiple Input-Queueing in ATM Switches", *Broadband Communications*, L. Mason and A. Casaca (eds.), pp. 382-393, Chapman & Hall, London, U.K. (1996).

[4] C. Kolias and L. Kleinrock, " Performance Analysis of Variations on Input-Queueing for ATM Switches", submitted.

[5] C. Kolias and L. Kleinrock, "Yet Another Analytic Study of Input-buffered ATM Switches", *Technical Report*, Computer Science Department, UCLA, Mar. 1997. http://millennium.cs.ucla.edu/~ck/tr1.ps

[6] C. Kolias and L. Kleinrock, "Performance Analysis of Multiplane, Nonblocking ATM Switches", *in Proc. IEEE Globecom '98*, Syndey, Australia, November 1998.

[7] L. Kleinrock, "On Flow Control in Computer Networks", in the proceedings of the in *Proc. IEEE ICC '78*, Toronto, Canada, pp. 27.2.1-27.2.5, June 1978.

[8] L. Kleinrock, "Power and Deterministic Rules of Thumb for Probabilistic Problems in Computer Communications", *Proc. IEEE ICC '79*, Boston, Massachusetts, pp. 43.1.1- 43.1.10, June 1979.

[9] H. Takagi, *Queueing Analysis, A Foundation of Performance Evaluation, Volume 3: Discrete-Time Systems*, Elsevier Science Publishers B.V., Amsterdam, The Netherlands (1993).