Five equations relating throughput capacity to system resources and risk for all agent-directed non-growth systems

James Bradley
Department of Computer Science
University of Calgary, Calgary, Alberta, Canada

Abstract Five equations for system throughput capacity (I), governing all non-growth, non-evolving, agent-directed systems are proposed and justified. Each equation covers a specific system aspect. Any two or more of the equations can be combined.

The equations are: A *resource-sharing equation* that shows how I can be maintained by reducing resources and increasing resource-sharing procedure complexity, or vice versa. A *basic risk equation* that shows how expected I increases [decreases] linearly with positive [negative] risk of loss of I in efficient environments. A *preventive-resources risk equation* that shows how I is improved by application of risk-preventing resources to reduce known risk. A *precautionary-procedure risk equation* that shows how I is improved by use of precautionary procedures to reduce known risk. A *monitoring-procedure risk equation* that shows how I is improved by use of a real-time monitoring procedure and risk-meaningful database to detect unknown risk and reduce it with precautionary response procedures. The conventional *standard deviation risk measure with respect to mean* from financial systems may be used, but a proposed new measure, called *the mean-expected loss measure with respect to hazard-free case*, is shown to be more appropriate for systems in general. The concept of an *efficient environment* is also proposed.

All quantities used in the equations are precisely defined and their units specified. The equations reduce to numerical expressions, and can be subjected to experimental test. The equations clarify and quantify basic principles, enabling designers and operators of systems to reason correctly about systems in complex situations. Spreng's Triangle, relating energy, time and information follows from the sharing equation. The empirical Markowitz-Sharpe-Lintner relationship between return, capital resources and risk for financial systems follows from the basic risk equation. The equations allow for the concepts of resources and time *margins of safety.*
Key words *Database, complexity, constraint, monitoring procedure, precautionary procedure, resources, resource-sharing procedure, risk, risk measure, throughput capacity.*

Introduction

In recent years the complexity of systems, particularly computer systems or systems with computer systems as major subsystems, has very greatly increased. As well, complex systems are increasingly having to cope with risk in their operating environment. Some systems are exposed to risk whose source is intrinsic, for example risk of deadlock in computer operating systems. Other systems deal with risk whose source is external to the system, such as an on-board aircraft computer system managing risk of colliding with a mountain.

Such systems are all human-agent directed systems. Unfortunately, the responsible agents may not always be able to reason clearly about the relationship between system throughput capacity, resources, resource-sharing procedures, environment risk, and risk-reducing resources and procedures. As a result serious mistakes in system design and system operation can occur, and particularly with systems whose throughput is human, not infrequently throughput loss means loss of human life.

In this paper we present five fundamental systems equations that relate system throughput capacity to resources, risk, complex risk-management and resource-sharing procedures, and which govern all non-growth, non-evolving, agent-directed systems. Agent-directed systems are

systems designed, constructed and operated by human agents; thus we exclude naturally occurring systems such as meteorological systems and living organisms. Non-growth, non-evolving systems are systems where the output is not fed back to alter, usually grow, the resources of the system; we therefore also exclude, for example, financial systems where output, such as interest payments, is added to the principal to generate an even larger output; and we exclude all living organisms, whose output is used to both grow and reproduce. However this still leaves a very large class of systems to which the proposed equations apply. The five equations apply as well to a chemical plant producing a polymer, as they do to a transportation system moving people or freight from input locations to output locations, or to a computer system converting input data to useful information, or retrieving data from a file or database. However the research leading to the five equations was inspired by phenomena in complex computer systems, and it is expected that it is on the computer systems arena that the equations will throw the most light. Nevertheless the five equations express a unity of principle in a great diversity of system phenomena. [The equations also appear to be valid for security systems or security subsystems, including computer security subsystems, since a system exposed to security violation is exposed to risk of throughput capacity loss. However, the equations were not developed with security systems in mind, nor does the author have any significant expertise in this large field of human endeavor. Accordingly no reference to security systems is made in the body of the paper, and the author can merely hope that readers who are security systems experts will find some utility in the equations.]

Each of the five expressions covers different system aspects, but any or all of them can be combined to cover complex system circumstances. Although the five expressions can be used for precise numerical analysis, it is important to grasp that their main virtue is that they clarify and quantify basic system principles, and so enable clarity of reasoning about specific costs and benefits of different system approaches and different methods of system operation, and so enable avoidance of serious error.

Mindful of the fact that most discussions of systems are weakened by lack of precise definitions and units of measurement, and of Kelvin' famous dictum "If what we are talking about cannot be measured and expressed in numbers, then our knowledge is of a very mean and meager kind", in this paper every attempt has been made to ensure that concepts are clearly defined, and that symbols appearing in the five equations denote quantities that are measurable and expressible in clearly defined units.

1.0 Overview of the systems equations

A system is considered to be any entity that converts inputs to outputs, and which may be composed of subsystems, connected either in series or in parallel, each of which is also a system. We also consider a system to be an entity that functions under the direction of one or more agents, usually human, and which employs essential resources R, often portable, in an environment E, enabling it to convert a set of inputs N, informational or material or both, to a set of outputs U per unit time, so that U is the system throughput. A system environment will be defined later. If the maximum value for U is I, then I is said to be the throughput capacity of the system, measured as the units of output per unit time when the system is operating at maximum capacity. In general, if the system is operating at a fraction p of throughput capacity I, then the throughput is $U = pI$.

In some cases there will be more than one type of throughput product, in which case we can declare throughput capacity of one of them as being the main throughput product capacity I, in units of product per time period, and assume, as will normally be the case, that each other product throughput capacity is a byproduct capacity that is a function of I, in most cases a linear function. Accordingly, in this paper, we deal only with I, considered to be either the only or the main throughput product capacity.

[Agent-directed system outputs are normally of greater value to the agents than the system inputs, so that is there is a net positive value v to the agent, and measurable by the agent, associated with system operation; this value v will, usually, to a fair approximation, be proportional to the throughput U according to:

$$v = ku - C$$

where k and C are approximately constant, C being the fixed costs of operating the system. If the system operating at maximum throughput capacity, this would mean that the maximum value, or value generation capacity, V is

$$V = kI - C$$

These last value expression should be taken a guideline only, since k and C are not true constants and are affected by economic considerations beyond the scope of this paper. However, for the sake of completeness we will occasionally make use of this expression alongside expressions for throughput capacity I, but only to indicate how system factors that do not affect I may be important from an economic viewpoint because of a potentially large effect on V.]

A reference summary of the five expressions for throughput capacity I follows. Minimal explanation is included here. There is a major section for each of the equations in the remainder of the paper.

1. The resource sharing equation:

$$I = KR(1 - T_S/T)[1 + sF_T(T_S)] \qquad (1A)$$

Independent variables under the control of the agent are R and T_S. K is a constant and R measures resources available to the system, alterable only in valid units of the type comprising R. T is a constant and measures the time for which I is computed. T_S has units of time per (basic harmonic) resource unit, and measures the execution time, and thus resource-sharing complexity, of a *normally complex, coordinated, resource-sharing procedure* for sharing resources within R. $F_T(T_S)$ is a growth function that has value 0 when T_S is zero and increases at an decreasing rate with increasing T_S to saturate at 1.0. The constant s is the *available sharing-enhancement potential*, where sR is the effective increase in R due to execution of a coordinated resource sharing procedure sufficiently comprehensive to saturate $F_T(T_S)$ at 1.0. T_S/T is a measure of the resource capacity diverted from normal operations to carrying out the resource sharing procedure, and is thus a measure of the negative impact of the sharing procedure on throughput capacity I. When there is no resource sharing, T_S and $F_T(T_S)$ are zero, and the expression reduces to I = KR. There is a value for T_S at which I is maximized, found by solving $dI/dT_S = 0$.

2. The basic risk equation

$$I = R[K + (b_{pb} - 1)r(E)] = RK + Rb_{pb}r(E) - Rr(E) = R[K + b_p r(E)] \qquad (1Ba)$$ /* for positive risk */

$$I = R[K + (b_{nb} - 1)r(E)] = RK + Rb_{nb}r(E) - Rr(E) = R[K - b_n r(E)] \qquad (1Bb)$$ /* for negative risk */

Independent variables under the control of the agent are R and r(E). I is now a mean or expected throughput capacity. r(E) is risk measure per unit R; the risk r(E) is risk of loss of throughput capacity, and is a function of the environment E relative to the system; the risk can be varied for constant R, by varying the environment E relative to the system; b_{pb}, b_p, b_{nb}, and b_n are

constants. Positive risk is risk it can pay to take on average; negative risk is risk it can not pay to take on average. Allowed risk measures are statistical measures, for example standard deviation risk with respect to the mean I, or a new proposed measure, more suitable for systems in general, called mean expected loss (MEL) risk with respect to the hazard-free I. Positive and negative risk, using statistical risk measures, cannot occur together. The environments allowable to the system must be *efficient environments*, as explained later, otherwise b_{pb}, b_p, b_{nb}, and b_n are not constants. $Rb_{pb}r(E)$ is the gross extra throughput capacity generated by exposure to the risk if the hazard risked does not occur; Rr(E) is the average loss in throughput capacity due to the hazard actually occuring. The difference is the net extra throughput capacity added (if positive risk), on average, by exposure to the risk.

3. The preventive-resources risk equation

$I = R(1 - aP)[K + (b_{pb} - (1-N(P)))r(E)]$ (1Ca) /* positive risk and p = 0 */
$= R(1 - aP)[K + (b_{pb} - (1-N(P(1-p) + pP/r(E))))r(E)]$ /* for coupling factor p > 0 */
$= R[K + (b_{pb} - (1-N(P)))r(E)]$ /* where a = 0 and p = 0 */

$I = R(1 - aP)[K + (b_{nb} - (1-N(P)))r(E)]$ (1Ca) /* negative risk and p = 0 */
$= R(1 - aP)[K + (b_{nb} - (1-N(P(1-p) + pP/r(E))))r(E)]$ /* for coupling factor p > 0 */
$= R[K + (b_{nb}- (1-N(P)))r(E)]$ /* where a = 0 and p = 0 */

Independent variables under the control of the agent are R, r(E) and P. The quantity P measures *preventive resources*, applied to prevent average losses Rr(E) due to known existing risk r(E). If P is large enough for N(P) to reach 1.0, P can completely eliminate the risk. N(P) is a growth function with value zero when P is zero, growing at a decreasing rate with increasing P to saturate at 1.0 The preventive resources P can be applied to either positive or negative risk. The environments allowable to the system must be efficient, else b_{pb}, b_{nb} are not constants. p is a risk-coupling factor with constant value between 0 and 1.0, to allow for possible coupling of P with r(E), such that, for $p > 0$, the effectiveness of P in reducing risk is coupled inversely to r(E), so that N(P) is a rising function of both P and 1/r(E), with N=0 for P= 0, and saturating at N=1 for large P and large P/r(E). The quantity aP is normally a small fraction of one, with a being a positive constant. However the constant a may be zero, and occasionally may be negative; the quantity aP is relevant only for systems where the presence of P also affects throughput capacity somewhat, independently of its beneficial effect of reducing risk. Where aP is non zero, there is a value for P at which I is maximized, found by solving $dI/dP = 0$.

4. The precautionary-procedure risk equation

$I = R(1- t/T)[K + (b_{pb}- (1-H(t)))r(E)]$ (1Da) /* for positive risk and p = 0 */
$= R(1- t/T)[K + (b_{pb}- (1-H(t(1-p) + pt/r(E))))r(E)]$ /* for coupling factor p > 0 */

$I = R(1- t/T)[K + (b_{nb}- (1-H(t)))r(E)]$ (1Da) /* for negative risk and p = 0 */
$= R(1- t/T)[K + (b_{nb}- (1-H(t(1-p) + pt/r(E))))r(E)]$ /* for coupling factor p > 0 */

Independent variables under the control of the agent are R, r(E) and t. T is a constant and measures the time period for which I is computed. t is the time per unit R taken to execute a precautionary procedure, which, for H(t) = 1.0, can completely eliminate losses due to the risk

r(E) *whose existence is known in advance;* t is a measure of precautionary-procedure complexity. H(t) is a growth function with value zero when t is zero; H(t) increases at a decreasing rate with increasing t and saturates at 1.0. The precautionary procedure can be applied to either positive or negative risk. The term KRt/T measures the negative impact on I of using the precautionary procedure. Allowable environments must be efficient, else b_{pb} and b_{nb} are not constants. p is a risk-coupling factor with constant value between 0 and 1.0, to allow for possible coupling of t with r(E), such that, for $p > 0$, the effectiveness of t in reducing risk is coupled inversely to r(E), so that H(t) is a rising function of both t and 1/r(E), with H=0 for t= 0, and saturating at H=1 for large t and large t/r(E). There is a value for t at which I is maximized, found by solving $dI/dt = 0$.

5. The monitoring-procedure risk equation

$$I = R(1 - t/T)[K + (b_{pb} - (1 - M(c)H(t/M(c))))r(E)] \quad (1Ea) \quad /* \text{ for positive risk and } p = 0 */$$
$$= R[K + (b_{pb} - (1 - M(c))r(E)] \quad /* \text{ for negligible t with } H(t/M(c)) = 1.0 */$$
$$= R(1 - t/T)[K + (b_{pb} - (1 - M(c)H(t(1-p)/M(c) + pt/M(c)r(E))))r(E)] \quad /* \text{ for } p > 0 */$$

$$I = R(1 - t/T)[K - (b_{nb} - (1 - M(c)H(t/M(c)))r(E)] \quad (1Eb) \quad /* \text{ for negative risk and } p = 0 */$$
$$= R[K - (b_{nb} - (1 - M(c)))r(E)] \quad /* \text{ for negligible t with } H(t/M(c)) = 1.0 */$$
$$= R(1 - t/T)[K + (b_{nb} - (1 - M(c)H(t(1-p)/M(c) + pt/M(c)r(E))))r(E)] \quad /* \text{ for } p > 0 */$$

Independent variables under the control of the agent are R, r(E), c and t. The variable c is a measure, in terms of unfolding environment regularity and a risk-meaningful database, of the total number of, or level, of constraints used in, and thus complexity of, a risk monitoring and detection procedure. This monitoring procedure operates in real time, continuously monitoring the unfolding environment, and uses constraint violations to alert the system in advance of the existence of risk, *whose existence is not known in advance,* in time to execute a precautionary procedure to avoid it, and taking time t per unit R. M(c) is a growth function with value zero when c is zero; M(c) increases at a decreasing rate with increasing c and saturates at 1.0. H(t/M(c)) is also a growth function going from 0 to 1.0 with increasing t for constant M(c), and with increasing t/M(c). A risk monitoring procedure can be applied to either positive or negative risk r(E). Allowable environments must be efficient, else b_{pb} and b_{nb} are not constants. H(t/M(c)) is also implicitly a function of r(E) and falls with increasing r(E). If the precautionary procedure is fully effective in a negligible time, then H(t/M(c)) is 1.0 with t close to 0 and the equations simplify. p is a risk-coupling factor with constant value between 0 and 1.0, to allow for possible coupling of t with r(E), such that, for $p > 0$, the effectiveness of t in reducing risk is coupled inversely to r(E), so that H(t/M(c)) is a rising function of both t/M(c) and 1/r(E), with H=0 for t= 0, and saturating at H=1 for large t, large t/M(c) and large t/M(c)r(E). There is a value for t at which I is maximized, found by solving $dI/dt = 0$.

6. Combinations of the five system equations

The sharing equation (1A) essentially enhances the system resources R to $R(1 - T_S/T)[1 + sF_T(T_S)]$, so that the sharing equation can be combined with any of the others by replacing R by the sharing enhanced value, as in:

$$I = R(1 - T_S/T)[1 + sF_T(T_S)]\,[K + (b_{pb} - 1)r(E)]$$

the combination with the basic risk equation.

Each of the four risk equations may be combined with one or more of the others. However, because of the fact that, for statistical reasons (destructive interference), risks are not

normally additive, so that naïve algebraic addition of r(E) quantities is sometimes invalid, the exact method of combination depends on circumstances. Example are discussed later at various points in the paper. Many complex combination circumstances are possible, and space in this paper does not permit a full discussion of this topic. There are really many possible combination circumstances, and many of them have yet to be researched in detail.

Measure of resources

The units of both R and P are units of anything of value to a human agent. Thus R could be measured in dollars, marks, printers or printer equivalents, microprocessors or microprocessor equivalents, or even, where the resources are mostly human, such persons as programmers. Also, since a resource can be valued by the human effort or work, or work equivalent, required to produce it, and since work is a measure of energy, then ultimately resources R can be measured in units of energy [32].

2. The resource sharing equation

In this section we demonstrate the veracity of the resource sharing expression:

$$I = KR(1 - T_S/T)[1 + sF_T(T_S)] \qquad (1A)$$

$$I = KR \qquad (1Aa) \quad \text{if } T_S = 0$$

When no coordinated resource-sharing procedure is involved, and no risk is present, coordinated resource-sharing time T_S is zero, making $F_T(T_S)$ zero, so that expression (1A) simplifies to $I = KR$. This expression states that, if we increase resources R (in a valid manner), then I will increase linearly with R. For example, if we construct a system that is an m-fold replica of the original system, with resources mR, then throughput capacity will be mI. We take $I = KR$ as axiomatic for valid changes to R.

The basic linear relationship between throughput capacity and resources

Suppose we construct the smallest possible miniaturization of the original system with resources R/n and throughput I/n, such that, when this smallest miniaturization system is replicated n-fold, we recover the original system with resources R and throughput I. We call this smallest possible miniaturization of the system the basic harmonic (system) of the original system. In addition, we call the original system the nth harmonic (replication) system, so that, for an nth harmonic system with resources R, the only valid increase or decreases in R in equation (1Aa) can be in multiples of R/n of the same type as R, or, more loosely, in basic harmonic resource units of the type already constituting R. We also call the resources R/n the basic harmonic resource unit of the system.

For a simple system consisting of 10 processors producing 30 x-units per week, increasing R by 5 similar processors will increase x-unit throughput capacity by 15 x-units; the basic harmonic resource unit is 1 processor, so that R can be changed in units of processors. However, changing R by adding resource units of a type different or not equivalent to those already constituting R (e.g. by adding 5 processors each capable of generating 6 x-units per week) is invalid as far as (1Aa) is concerned.

The importance of expressions (1Ac) holding only for R being alterable in terms of basic harmonic resource units of the type already constituting R can be can perhaps be more forcefully illustrated by the following less obvious example. Suppose a software house with 4 programming teams, each team being 2 cooperating programmers, so that R = 8, when measured

in programmers. Suppose also that each team functions independently to produce x average application programs per year, for throughput capacity I of 4x programs for the house. Thus the basic harmonic resource unit is a team of 2 programmers. If we add 2 independent programmer pairs or teams to the original 4, so that R = 12, I goes to 6x; the increase was valid, being 2 basic harmonic resource units. But suppose we add 4 programmers by adding one programmer to each of the former teams, so that we have 4 new teams, each of 3 cooperating programmers; although R now also measures 12, this is not an allowable alteration of R, for we do not have an increase in independent resource units of the type already in R, that is, an increase in the number of basic harmonic resource units. Instead we have created a new system, and in all likelihood, in this case, new capacity I will not be 6x, but less, or maybe even I < 4x!

Parallel and serial subsystems

The restriction on the valid variability of R, in I = KR, to changes in harmonic resource units of the same type as R, has significant consequences with composite systems consisting of either parallel or serial subsystems. We consider parallel subsystems first.

A. Parallel operation

Suppose a system consists of two types of resources, say i low capacity processors amounting to resources R_1 (e.g. 15 slow processors that each generate 3 x-units per week per processor) and j high capacity processors amounting to resources R_2 (e.g. 10 fast processors that each generate 6 x-units per week per processor) with i > j. [We are not interested in economics here, just in throughput capacity – the upkeep cost of the slow processors may be less than half that of the fast processors, and so be more economic.] Suppose R_1 and R_2 operate in parallel, generating throughput I_1 x-units per time unit and I_2 x-units per time unit respectively, where the number of R_1 units can be increased or decreased independently of the number of R_2 units, and vice versa. In such a case the expression

$$I = K(R_1 + R_2) \qquad (1a)$$

will not hold for independent variability of R_2 and R_2. We have essentially two separate systems, or parallel subsystems, so that $I_1 = K_1R_1$ and $I_2 = K_2R_2$, and

$$I = I_1 + I_2 = K_1R_1 + K_2R_2 \qquad (1b)$$

$$[= 3R_1 + 6R_2 \text{ x-units per week, in the processor example}]$$

for the composite system, where R_1 is alterable in one set of independent units (slow processors), and R_2 in another (fast processors). The expression $I = K(R_1 + R_2)$ will hold only if, when R_2 is increased or decreased by R_2/n units, R_1 is increased or decreased by R_1/n units where n is the largest common divisor of R_1 and R_2, that is, if we increase in sets of R_1 and R_2 units (that is, in basic harmonic resource units, of sets of 3 slow and 2 fast processors, for $I = (21/5)(R_1 + R_2)$, giving an nth harmonic system for n = 5.). [If n is very small, for example, n = 1 with 10 slow and 7 fast processors, we may prefer to perform a simplifying adjustment to the system, in the interests of a tidier, or more finely-grained, harmonic system: either by adding 4 slow processors (n = 7) or 3 fast processors (n = 10).]

B. Serial operation

If two subsystems each with resources R_1 and R_2 are operating in series, everything depends on whether of not both subsystems are operating at capacity, and to what extent the subsystem resources can be shared.

If both subsystems are operating at capacity, then throughput capacity $I = K(R_1 + R_2)$ holds. If it is a nth harmonic system, a valid increase is an increase in units of $(R_1 + R_1)/n$. However, if R_2 is operating below capacity, and R_1 is at capacity, then R_1 is the *limiting (or "bottleneck")*

resource, and R_2 is the *non-limiting resource*. Although $I = K(R_1 + R_2)$ holds, it is also the case that $I = K_aR_1$ holds for increases in R_1 up to the point where R_2 starts to operate at full capacity, at which point $I = K_b(R_1 + R_2)$ holds for further increases in resources. [The concept of a non limiting resource is needed later to help define a system environment.]

Coordinated and non-coordinated sharing of limiting resources

Suppose again that two subsystems each with resources R_1 and R_2 are operating in series. If one of the resources R1 is limiting then R1 may be capable of participating in either *uncoordinated exclusive-allocation sharing*, or *coordinated inclusive-allocation sharing*. In both cases there is sharing, but in each case the nature of the sharing is very different, and the difference is vital for understanding both equation (1A) and the operation of complex systems. Rather than define the two resource-sharing concepts above at this stage, three quite different instances are given below, in order of increasing complexity, to promote the reader's understanding of the differences between them.

Instance 1: Railroad Example Consider a simple railroad system with a single track (R_1 resources) connecting A and B, and 2 physically different trains (R_2 resources). It takes 3 hours from A to B and 3 back again, with 9 hours loading and unloading time at each of A and B. This means 2 trains per day, with capacity being 4 trainloads per day. At this point R_2 is limiting, and capacity increases linearly with R_2 until R_2 reaches 4 trains, in accordance with $I = KR_2$, at which point we have 8 trainloads per day and now R_1 is also limiting. At this point, during each time unit (1 day) the track is shared between 8 trainloads, but at any given instant there is only 1 trainload out on the track. The track is exclusively allocated to one trainload of throughput at any instant. If we increase the number of physical trains to 8, and retain exclusive track allocation, throughput capacity remains at 8 trainloads per day, and either 4 trains never run or each train runs every other day.

We can increase throughput capacity however, if we abandon exclusive allocation of the track and allow more than one trainload on the track at once, that is, we share the track with inclusive allocation. This means that trains will have to pass, and thus will have to be coordinated, with one train stopped at a siding, which reduces the time during which the track is available for traffic, while the other train passes. Thus if we have two trains on a track at a time we can make use of 8 physically different trains each day, for capacity of 16 trainloads per day; if we can have 4 on a track at a time we can use 16 trains each day and have capacity of 32 trainloads per day. This is an example of coordinated sharing. As the number of trainloads sharing the track increases the amount of coordination time T_S, during which there is reduced availability of the track for traffic, must increase. There will normally be some maximum number of trains that can be on the track at any instant, determined by the number and location of passing areas. Equation 1A covers this situation.

Instance 2. Visiting salespersons Consider a sales company with a passenger van (R_1) and 2 salespersons (R_2) and an 8-hour workday, where throughput capacity is measured in the number of visits by salespersons to client sites per day. Suppose a visit takes 30 minutes and for every 2 visits 3 hours of office desk work by the salesperson at the company office is needed. Clearly, capacity is 8 visits per day and the van is not limiting. As R_2 is increased, throughput capacity increases to 16 visits per day when R_2 is 4 persons, and to 32 visits per day when there are 8 salespersons, in accordance with $I = KR_2$, at which point the van R_1 is now limiting, under exclusive allocation. For each visit a salesperson takes the van, and thus has exclusive control. The van is shared, but the sharing is uncoordinated. If we now double R_2, throughput capacity

remains the same, and either some sales persons never make a visit or each makes only 2 visits per day instead of the 4 each is capable of.

We can increase throughput capacity if we abandon exclusive allocation of the van and share it among visits, that is, visits could be coordinated so that, more or less, via short detours and drop-offs and pick-ups, the van can allow more than one person to make two visits in the same hour, that is, the van is shared among visits by means of a coordinated sharing procedure. This type of sharing, unlike the previous uncoordinated type, will absorb some of the van's time, because of the coordination time T_S taken up by detours, drop-offs and pickups, so that the time during which the resource is actually available for visits is reduced, but because of the increased level of sharing, throughput will normally be increased, unless the detour/drop-off/pick-up T_S time is excessive. This situation is covered by equation (1A)

Instance 3. Computer Operating System. Consider a computer system with a fast processor/memory unit (R_1) and 2 slow input-output device pairs (R_2). Each I/O device pair is at different user locations, and is in use for 10 minutes in each hour, and during the remaining 50 minutes of the hour new data is being prepared for entry. During use of an I/O pair, input data (i.e., a "jobload" of data) is interactively entered and converted via processing to information output on the output device [13]. During a 10 minute session, the processor is exclusively allocated to a specific I/O pair. Clearly, throughput capacity is 2 jobloads per hour. As we increase the number of I/O device pairs, capacity increases, in accordance with $I = KR_2$, to 4 jobloads per hour for 4 I/O device pairs, until finally with 6 I/O pairs capacity is 6 per hour. At this point the cpu/memory unit R_1 becomes limiting. R_1 is being shared in an uncoordinated manner among the jobloads, with exclusive allocation of R_1 to each jobload, that is, a jobload must finish being processed before the processing of a new one can begin. If we now double the number of I/O device pairs to 12, further throughput capacity increase will not occur, but either 6 device I/O pairs will never be in use, or each I/O pair will be operating at half capacity, once every second hour, assuming we retain exclusive allocation of R_1 to a jobload.

Suppose the processing of a jobload, as is usually the case, consists of short cpu processing bursts. Suppose that each burst lasts 60 milliseconds on average, and 100 bursts are needed to process a jobload, so that 6 seconds of actual processing time are needed per jobload of data. In that case, by interleaving different jobs [1, 19], that is, by coordinated sharing of R_1 inclusively among multiple jobloads, we could process up to 594 extra jobloads per hour on top of the 6 that renders R_1 limiting. This increase is spectacular (because of the enormous cpu speed), compared to the increase in throughput capacity that can normally be obtained by coordinated sharing with inclusive allocation with other types of systems. Nevertheless, although less obvious in this case, the principle is the same. At any instant the (limiting) processor/memory unit R_1 is being shared among the processing of more than one jobload (actually among 100 jobloads). In practice, given the numbers above, somewhat less than an extra 594 jobs will be achieved by the coordinated sharing, since the R_1 will be devoted to coordinated sharing activities (processor scheduling, context switching, dispatching, etc.) during a time T_S. Equation 1A applies to this situation.

Units of the rate constant K and throughput capacity I

In the expression $I = KR$, the constant K is called the rate constant. Its units can be either intradenominational or interdenominational, intradenominational units occurring mostly in financial systems. Consider R = \$1,000 invested in bonds at an interest rate of 10% per annum; this constitutes a simple financial system, generating throughput I of \$100 per year, so that $I = KR = 0.1R$, with the rate constant K being the intradenominational per-unit interest rate, being measured in dollars per unit dollar per year. [When measured as dollars per 100 dollars, it is the

percentage interest rate.] Now suppose that $1 = 1000 lira, and that the lira interest rate is 5%. Now if R is in liras but interest is paid in dollars we have I = 0.000005R, for an interdenominational (per-unit interest) rate of 0.000005 dollars per lira per annum. Conversely, if R is in dollars but I is in liras, the interdenominational return rate K is 100 liras per dollar per annum.

With non financial systems K is nearly always interdenominational, being expressed in such units as eggs per hen per week, or jobs per processor per minute, or wheat bushels per 20-ton truck per year, or bytes per second per disk-drive controller. Such units are in common everyday use in the practical business of production. Note however, that it is possible to have intradenominational units for K in non financial systems if we measure R in dollars or joules, and I in dollars or joules per time period, so that K is dollars per dollar per time period or joules per joule per time period. This will be relevant later, when Spreng's triangle is considered.

In general K is measured in units of output per time period per harmonic resources unit. In the simpler cases a harmonic resource unit is 1 processor, or 1 20-ton truck or 1 cpu. However, in more complex cases a harmonic resource unit will be made up of combinations of units of different resources, so that we can have K expressed as widgets per week per (harmonic resources) set of 2 lathes and 1 operator and 3 computers. However, in cases where such an obviously awkward unit is nevertheless the correct one, an improvement may be to use the aggregate of a common attribute type, such as dollar value, area, weight, volume, as in eggs per acre of (basic harmonic unit of) system, or megabytes per cubic meter of (basic harmonic unit of) system, and so on.

Throughput capacity I can be measured in throughput entity or throughput entity set units, or in throughput entity attribute units. When we use throughput entity units we use the number of entities throughput per time period, e.g. eggs per week, autos per day, passengers per year, computer jobs per second, files per hour, pages per second, and so on. With entity set units, we simply use named sets of entities throughput per time period, e.g. crates (of eggs) per week, crates (of onions) per year, truckloads (of bricks) per months, containers (of apples) per year, trainloads of wheat per week, diskloads of data per day, and so on. When we use attribute units we use some measurable attribute of the entities, which relates linearly to entity set units, but which is common to other entity types as well, for example, tons of eggs per day, tons of apples per week, cubic meters of oil per day, cubic meters of water per hour, and so on.

The concept of coordinated sharing of a resource

The entity or entity set per time period measure of throughput enables a concise definition of coordinated sharing of a resource:

Resources R of a system or subsystem are being shared in a coordinated manner if, and only if, at any instant, more than one entity or entity set unit of system throughput is under processing by resources R, requiring that the R resources be engaged in (usually complex) coordination procedures for measurable periods of time.

Thus a single railroad track is being shared in a coordinated manner if it is handling multiple trainloads of entities at any one instant. An automobile with two occupants is being shared in coordinated manner if each occupant has a unique list of places to visit, and at any instant the car is engaged in enabling a visit for each person. A computer processor or human is participating in coordinated sharing if it (he/she) is handling more than one job or process at any instant.

Relationship between I, R and coordinated sharing time (or complexity) T_S

Consider any basic system of throughput capacity I and resources R, of arbitrary complexity, where some subsystem resources are connected in series and others in parallel, and

some are shareable and some not. It follows from the discussion earlier that when the system is operating at capacity, some subsystems will be at capacity and others below capacity.

We can define an *available coordinated sharing-enhancement potential* s per harmonic resource unit of the system, where $s > 0$, such that resources R are effectively increased by sR, and thus throughput capacity I is increased by sKR, by means of coordinated resource-sharing up to the physical limit possible, so that after such resource sharing, $I = KR(1+s)$. The value for s for a given system can only be obtained by analysis of the system. For example, in the case of the computer system earlier with the processor/memory unit limiting, R can be taken as 1, and K as 6, so that $I = KR = 6*1$ with no coordinated sharing. In theory, with coordinated sharing up to the limit possible, we have $I = 6R(1 + 594/6) = 100$ jobloads per hour, so that $s = 594/6 = 99$.

If only a fraction F of the available sharing-enhancement potential s is made available through limited coordinated sharing, I will be less than its potential $KR(1 + s)$ and will be given by $I = KR(1 + sF)$. F might be called the *available sharing-enhancement potential fraction.*

With any system, coordinated resource sharing among throughput entities requires that system resources R participate in a usually complex and time-consuming *coordinated resource-sharing procedure* to enable concurrent handling of multiple throughput entities without collisions. An important property of the sharing procedure is that is it requires allocation of resources R being shared for total time T_S of operation of the sharing procedure (per basic harmonic resource unit), thus temporarily diverting the resources from normal throughput generation. System agents, while seeking the increased throughput capacity benefits of resource sharing, will be inclined to seek to minimize system time T_S per harmonic resources unit lost to operation of the sharing procedure.

We now assert, as axiomatic, that, generally, system agents will undertake coordinated resource-sharing in steps, where the most effective, in terms of most throughput capacity increase for least sharing time T_S, is undertaken first, followed by the next most effective, and so on. Hence, as the fraction F of the available sharing-enhancement potential s, due to operation of a coordinated sharing procedure, increases towards 100%, each increased 1% gain in F requires a greater increase in coordinated resource-sharing procedure activity time T_S than for the previous 1%. This means that the available sharing-enhancement potential fraction F must be a function of T_S, that is $F = F_T(T_S)$. Empirically, $F_T(T_S)$ must have value 0 when T_S is zero and must increase at a decreasing rate with increasing T_S, to saturate at value 1.0.

Hence, if we neglect the costs of resource sharing, due to resource diversion for time T_S during operation of the sharing procedure, the relationship between coordinated resource-sharing procedure time T_S per harmonic resource unit and throughput capacity I must be given by an expression

$$I = KR(1 + sF_T(T_S))$$

where $F_T(x)$ is a growth function of the general form:

$$G(x) = (1 - e^{(-x/g)})$$

The constant g can be very small, so that G(x) quickly becomes 1.0 when x is only slightly greater than 0, corresponding to the case where very little resource sharing time T_S is needed to achieve 100% of available sharing-enhancement potential s. Note, however, that since $F_T(T_S)$ is essentially an empirical expression that will represent some average of a large number of similar functions for different situations, some departure from the ideal G(x) shape can be expected in any practical situation, although in all cases it will increase on average from 0 with increasing T_S to saturate near 1.0. Since T_S measures the time for which the sharing procedure is in operation, per basic harmonic resources unit, it may also be regarded as a measure of the complexity of the coordinated sharing operation, and thus the order created [31] by the sharing operation, and hence the decrease in entropy of the system.

In the above expression we have neglected the negative impact on I of operating the coordinated resource-sharing procedure during time T_S. If T is the period during which throughput capacity is measured, for example T = 1 hour, where I is in jobloads per hour, or T = 1 day, where I is trainloads per day, then the fraction of the time during which R is diverted from normal throughput is T_S/T. Hence operation of the sharing procedure effectively reduces R to $R(1 - T_S/T)$. Hence the correct expression for the relationship between I, R and T_S must be

$$I = KR(1 - T_S/T)[1+ sF_T(T_S)] \qquad (1A)$$

which is the resource sharing equation.

Equation (1Aa) is stating that following introduction of a coordinated resource sharing procedure taking time T_S per basic harmonic unit of resources, with $F_T(T_S)$ anywhere between 0 and 1.0, and consequent increase in I is as given in (1A), it is then possible to increase(decrease) R in basic harmonic units subjected to the same level of sharing, and obtain a further increase(decrease) in I. However, R has to be increased in valid units, which means in units of resources equal to the smallest functional miniaturization of the system resources, or basic harmonic resources units. Also the increases in R must be shared in the same way as exiting units of R. It is important to understand that T_S is the time taken for sharing the resources in such a minimum unit of resources, and is thus independent of the magnitude of R. The correct units of T_S are therefore units of sharing time per unit of basic harmonic resources. Of course, after some time, if the system is operated with a fixed or standard level of sharing, then $(1 - T_S /T)[1+ sF_T(T_S)]$ is constant, and can be absorbed into the rate constant K as K_S in $I = K_S R$, the resources of any increase in R being assumed to be shared in the standard manner.

It can happen that too much of the extra throughput benefits from sharing resources may be offset by the cost of the sharing, so that the criterion for productive use of a sharing procedure is obviously:

$$(1 - T_S/T)[1+ sF_T(T_S)] > 1$$

Where this inequality does not hold, coordinated sharing cannot pay.

If equation 1A is analyzed carefully, and examples pondered, it will be seen that there is a level of sharing for which I will be a maximum. If T_S is small relative to T when $N(T_S)$ approaches 1.0, then this maximum will be

$$I = KR(1 - T_S/T)[1+ s]$$

This reduces to $I = KR[1+ s]$ if T_S is really small compared to T as $N(T_S)$ approaches saturation.

At the opposite extreme, T_S could approach T before $N(T_S)$ reaches saturation, at which point I will approach zero. Thus in the general case I will increase with increasing T_S to a maximum value, and then decline, possibly to zero, as T_S continues to increase. A declining I for increasing T_S is the phenomenon of thrashing, where increasing the level of coordinated sharing results in less throughput capacity because the incremental cost of operating the sharing procedure exceeds the incremental benefit. Thrashing is a well-known phenomenon in computer operating systems [29], but is possible in any system with coordinated sharing. If we set $x = T_S$, the peak value for I is obtainable from a solution of

$$dI/dx = - KR/T[1+ s(1 - e^{(-x/g)})] + KR(1 - x/T)(1/g)se^{(-x/g)} = 0$$

which, since it involves standard calculus, is left to the reader.

It should now be clear that the increased throughput capacity caused by coordinated sharing of resources within a system always has to be "paid for" by means of resource-consuming complex coordinated resource-sharing procedure activity, in which the complexity may be measured by T_S, and this axiomatic rule cannot be evaded. Readers who merely look around, will see it in operation in almost every agent-directed system, whether simple or complex, whether technological or social

Spreng's Triangle

D. T. Spreng, a physicist, hypothesized that for any given task, for example to produce output q,

$$f_1(t)f_2(E)f_3(i) = \text{a constant}$$

where t is the time, E is the energy, and i is the information required to perform the task, and where $f_1(x)$, $f_2(x)$, $f_3(x)$ are functions whose value generally increases with increasing positive x. This is one statement of what has come to be known as Spreng's Triangle [2, 30]. In other words to carry out a given task, you can save energy by taking more time or using more information, or you can save time by using more energy, and so on.

Spreng's triangle can be deduced from the resource sharing equation:

$$I = KR(1 - T_S/T)[1 + sF_T(T_S)] \qquad (1Aa)$$

Suppose we need to generate a quantity of output q. If this is done in time t with the system operating at maximum capacity then q/t = I and:

$$q = KtR(1 - T_S/T)[1 + sF_T(T_S)]$$

In this expression, we can let q represent the fixed task, so that

$$tR(1 - T_S/T)[1 + sF_T(T_S)] = \text{a constant}$$

But R is a measure of the energy required in joules, so that $R = f_2(E)$, and T_S is a measure of the order in the sharing procedure, hence a measure of negative entropy, hence a measure of information. Hence $(1 - T_S/T)[1 + sF_T(T_S)]$ can be written as $f(T_S)$, which can be written as $f_3(i)$ where i is information. Hence:

$$tf_2(E)f_3(i) = \text{a constant}$$

It follows that the sharing equation is stating some fundamental limits about the nature of physical reality.

Resource R adjustments

As we have seen, the increase in throughput capacity due to coordinated sharing is caused by the resources of at least one limiting subsystem being shared among the throughput entities feeding to or from underutilized subsystems. However, as a result of the sharing, sometimes adjustments to R, not covered by equation (1a), are desirable. [Note that this subsection on adjustments is included for purposes of completeness, and may be skipped on initial reading.]

Take a system consisting of a computer processor/memory unit with many input and output device pairs with no coordinated sharing of the processor. A jobload of data coming from an input device (in continual disparate input bursts) and going to the output device of the I/O device pair (also in continual disparate bursts) has the processor exclusively allocated until the jobload of data has been completely output. The processor/memory unit is the limiting resource and so many I/O device pairs are underutilized. But if the processor can have multiple jobloads under processing at once, with the processing of the bursts of one jobload being interleaved among the bursts from another job, we can have sharing of the limiting resource, and because of higher throughput capacity, consistent with (1A), conversion of the I/O devices from underutilized to better or fully utilized resources.Following sharing of limiting resources to the limit possible, there are essentially two distinct possibilities for adjusting R by adding further subsystem resources to allow for better subsystem matching.

Case 1. Adjustment with unsharable resources R_u Suppose that the computer system with resources R has 1 processor (with sufficient memory) and 10 I/O device pairs, each pair being used for a single job or jobload of data. Now suppose that the processor is shared, in an inclusive coordinated manner, among the jobloads from all of the I/O device sets, so that $F(T_S) = 1.0$, but the situation is such that the processor could be shared among the jobloads from12 I/O device

pairs as a maximum. In that case an even greater sharing-enhancement of R is possible, that is, a larger s is possible. An adjustment of 2 sets of I/O-device unsharable resources R_u could be added to R, for a total of 12, such that if the sharing procedure and thus T_S is extended to cope with them, additional throughput will result, with a new version of (1A) being applicable:

$$I = K_u (R + R_U)(1 + T_S/T)[1+ s_u F_{Tu}(T_S)]$$

Addition of R_u to R is not a valid increase in R in (1A), and causes the original system to be fundamentally changed with respect to resources, so that modified constants K and s, and function $F_T(T_S)$ are now required.

Prior to undertaking additional sharing to cope with the addition of R_u, $F_{Tu}(T_S)$ will be less than 1.0. Adding appropriately to T_S will bring $F_{Tu}(T_S)$ to saturation at 1.0, thus completing sharing to the maximum possible allowed for by the nature of the processor. At this point resources $R + R_u$ may be said to constitute not only a basic harmonic resources unit, but a sharing-potential maximized basic harmonic resource unit. and from this point resources can be increased in units of $R + R_u$ with maximum efficiency resulting, that is, maximum increase in I per resources unit.

Case 2. Adjustments with sharable resources R_S Suppose that the computer system with resources R has 24 I/O device pairs, each pair being used for a single job or jobload of data. Now suppose that, because of the nature of the processor it can be shared among the jobloads from 12 I/O device sets at a maximum, so that $F(T_S) = 1.0$. This means there are 12 extra I/O device pairs than cannot be used efficiently and which are contributing nothing to capacity. But this does not mean that they can be removed from the system without effect, for they can still be in use! Every pair is likely to have users, and which pairs are regarded as superfluous as far as throughput capacity is concerned is arbitrary, for we can simply regard all the I/O device sets as being one half underutilized.

To deal with this situation by improving throughput capacity, we could adjust R by adding 1 processor (sharable resources R_S) to give 2 processors and 24 sets of I/O devices, so that a new version of (1A) is applicable:

$$I = K_g (R + R_S)(1 + T_S/T)[1+ s_g F_T(T_S)]$$

Addition of R_S to R is not a valid increase in R in (1A), and causes the original system to be fundamentally changed with respect to resources, so that modified constants K and s, but this time no modified function $F_T(T_S)$, are now required.

Prior to the adjustment, the sharing time T_S was for 1 processor being shared among the equivalent of jobloads from 12 sets of I/O devices. Following the adjustment, the basic harmonic resources unit is 1 processor and 12 sets of I/O devices, so that exactly the same sharing time T_S per basic harmonic resource unit is required, and thus the same function $F_T(T_S)$. At this point resources $(R + R_S)/2$, or 1 processor and 12 I/O device sets, are the basic harmonic resources unit, and also the sharing-potential maximized basic harmonic resource unit. From this point resources can be increased in units of $R + R_S$ with maximum efficiency resulting, that is, maximum increase in I per resources unit.

Note however, with R being 2 processors and 24 sets of I/O devices, if we continue to adjust R by adding processors, which are sharable resources, only minor and possibly no increased throughput capacity results. The level of coordinated sharing, however, as measured by the value $F_T(T_S)$, must fall, since we are increasing the number of processors per I/O device set, until when there are 24 processors there is no sharing at all, $F_T(T_S)$ is zero, and each set of I/O devices has its own processor. It is just such an adjustment that occurs when, in the example

earlier of visiting salespersons with only one shared van, they buy a fleet of vans, one for each salesperson, and eliminate the complexity of coordinated van sharing entirely.

Risk of capacity loss due to deadlocks and collisions

A second-order effect due to the use of a coordinated sharing procedure is that sometimes risk of throughput capacity loss appears as a side effect, for example, due to deadlocks [14, 17] and critical-section (interference or collisions) problems [18, 25] in computer operating systems. In general, such risks are the consequence of mismatches between system resources for a given sharing procedure and input stream, and will be considered later.

3.0 Risk versus resources

Risk of loss of anything of value is normally run to secure some gain in a value, and is therefore of primary interest to humans. Consequently, the concept of risk has been thoroughly studied in the financial industry. A major result is that we have a sound measure of risk, at least in a financial context [8, 10], that is, *the standard deviation risk measure with respect to the mean*, primarily due to Markowitz [24]. Risk is, however, known to be a slippery concept, and interested readers unfamiliar with the statistical concept of risk may need to become familiar with some of the literature on the subject. When systems in general are considered, we find that a new measure, somewhat different from the conventional standard deviation risk measure used in finance, is very useful, as will be developed shortly,

One key to understanding risk is grasping the distinction between exposure to *the certainty* of future loss and exposure to merely *the possibility* of future loss. In both cases there is exposure to future loss. But *only exposure to the possibility of future loss is risk.* In some cases there is exposure to future loss consisting of both exposure to the certainty of future loss and to the possibility of an additional future loss, for example, if exposure to a future loss of 20 where a future loss of 16 is certain and a future loss of 4 is additionally possible, there is only risk of loss of 4. Thus there has to be a variability aspect to risk, since risk is only a possibility of future loss: in some future periods the loss will occur fully, in others partly and in others not at all. Failure to grasp these distinctions has caused many fruitless arguments and debates about correct measures of risk

In this paper we are primarily concerned with risk of loss of system throughput capacity I, the system being considered to be financial only if throughput capacity I is in actual currency entities, such as dollars, and not in other physical units, for example, digital documents, valued in currency units. But first we need to consider risk measures, both the conventional standard deviation measure from financial systems, and a new proposed mean-expected-loss measure for systems in general.

Risk measures

In general risk of loss of throughput capacity has two components, namely the probability of a hazard occurring and the size of the loss in throughput, with respect to some standard level, should the hazard occur. However, in a system situation where there is exposure to possible loss with respect to some standard level of throughput capacity, there will often be exposure to possible gains in addition, depending on the standard level used. An accurate risk measure must therefore combine these different aspects of risk, but must not include any measure of certain future loss.

Suppose the system is exposed to unpredictable losses and gains in throughput capacity, that the statistics of these fluctuations are constant (or stationary [4], in statistical terminology),

and that over a long enough period of time to be representative of these statistics, the mean, and thus expected, throughput capacity is I_m, and in n fully representative time periods the actual capacity values are:

$I_m - L_1, I_m - L_2, \ldots I_m - L_i, I_m + G_1, I_m + G_2, \ldots$ or $I_m + G_j$

where L_1, L_2, ... are deviations downward (losses) from the average I_m, and G_1, G_2, ... are deviations upward (gains) from I_m, with $n = i + j$, so that

$$(L_2 + L_2 + \ldots L_i) = (G_1 + G_2 + \ldots G_j)$$

The same throughput capacity deviations can then be expected to occur in the future in unpredictable order, all equally likely.

[Note that the above quantities should be interpreted as follows: Suppose n = 10, i = 6 and j = 4, and the losses L_1 .. L_i are 30, 20, 20 10, 10, 10, and the gains G_1 .. G_j are 40, 20, 20, and 20. Then imagine an urn containing 10 balls: 6 red balls each with one of the losses marked, and 4 green balls each with one of the gains marked, and thus the distribution of future (and past) losses and gains; to simulate what will happen in the next time period, select a ball randomly from the urn, the number on the ball giving the amount; to simulate for the subsequent time period the ball removed must first be replaced, i.e. we must use selection with replacement. The sum of the numbers on green balls divided by 10 is the average or expected gain, equal to the red ball sum divided by 10, the expected or average loss. Nevertheless, suppose the period involved was 1 week; then in some weeks there would be a gain, in others a loss, and in few weeks a severe loss or a large gain. Thus the observed result is merely a sequence of unpredictable losses and gains per time period, with respect to the mean, that is, I will fluctuate from one period to the next. In terms of probabilities, the risk is due to probabilities 0.1. 0.2, 0.3 of losses 30, 20 and 10 respectively, and probabilities 0.1, 0.3 of gains 40 and 20, all with repeat to the mean.]

The expected or average I_m actually rarely occurs if at all. In reality all we have is the unpredictable sequence of losses (L_1, or L_2, ...) and gains (G_1, or G_2, ...) with respect to an average or expected throughput capacity I_m in a given time unit, and it is such losses and gains with respect to expected throughput I_m that must be used in the measure of risk of loss of throughput.

For a meaningful measure of risk there are now two choices, the traditional standard deviation measure, and a new measure that in many cases is more suitable for systems in general.

Choice 1. Take the standard deviation of the deviations (L_1, L_2, ... G_1, G_2, ...) from the mean throughput capacity I_m, as *a standard deviation measure of possible loss with respect to (or down from) the mean I_m*, to give the *Standard Deviation (SD) risk* measure.

If we use twice the standard deviation we have an even stronger risk measure, the *2-Standard Deviations (2-SD)* risk measure.

Interpretation: A SD-risk of s means that in the next time unit, there is a 50% chance or possibility of a loss down from the expected I_m, and a 34.1% chance of a loss between 0 and s down from the expected I_m, and a 15.9% chance of a loss > s. In addition there a 47.7% chance of a loss between 0 and 2s with respect to the mean I_m. This implies, that there is a 13.6% chance of a loss between s and 2s, and a 2.3% chance of a loss >2s, both losses with reference to the mean throughput capacity I_m. In specifying an SD risk, *we must both specify the standard deviation and specify with respect to what standard level.* A 2-SD-risk of 2s means that in the next time unit, there is a 50% chance of a loss with respect to the expected I_m, and 47.7% chance of a loss between 0 and 2s and a 2.3% chance of a loss > 2s.

The percentages used are from a normal distribution function table, and assume that losses and gains in each time unit are distributed normally. The numbers needed are different if the distribution departs from normal.

[The SD-risk measure is the one widely used in finance, particularly for stock and bond portfolio management, for which it is both correct and adequate [8, 24], since stock and bond prices follow close to a random walk, which gives rise to a near-normal distribution of price changes [4]. Notice that where there is exposure to future loss, where the future loss includes a certain loss and a possible loss, the SD-risk measure selects out only the possible loss, that is, the true risk. For example, suppose a system where ideally I = 400 if there were no future loss exposure, but where actually the system has exposure to a future loss in I whose mean is 100 and whose standard deviation is 14, where the least loss is always greater than 70. That means a certain loss of 70 plus a loss whose mean is 30 that can be as small as 0 and as large as about 60, with a standard deviation of 14, that is, a certain loss of 70 plus a standard deviation of 14 about the mean of the loss variations of 30, that is, certain loss of 70 plus an SD-risk of 14 with respect to a mean of 300.]

However, to deal with the problems and possibilities in arbitrary systems, an additional and complementary risk measure is very useful. This is the MEL-risk measure defined below. [The reader who is expert in financial risk analysis using the SD-risk measure may be want to immediately dismiss this additional risk measure as nothing but an intellectual crutch; the author asks such readers to suspend judgement until after studying the use of the MEL-risk measure with preventive resources, precautionary procedures, and monitoring procedures, which are risk-measure applications not dealt with in conventional financial risk management.] The author therefore proposes:

Choice 2. Suppose that for a system exposed to risk, there is at least one hazard-free time period, in which by chance the hazard risked does not occur, and where the gain with respect to the mean throughput capacity I_m is G_b in this hazard free time period, and where a gain exceeding G_b is thus not possible (but a gain under G_b is possible), for a total hazard-free throughput capacity of $I_m + G_b$. Then all other throughput capacities $I_m - L_1$, $I_m - L_2$, ... , $I_m + G_1$, $I_m + G_2$, ..., each in a time period where the hazard does occur in varying degrees of intensity, may be considered as exhibiting losses, or loss deviations, $G_b + L_1$, ...$G_b - G_1$ ... down from, or with respect to, the value of I in the hazard-free time period. We may use the mean of these loss deviations (down) from I for the hazard-free time period as a measure of the risk, that is, a measure of expected losses in the future with respect to the throughput capacity for a hazard-free time period, that is, the *Mean Expected Loss (MEL) with respect to, or down from, the throughput capacity in any hazard-free time period,* or *MEL-risk.*

Note in specifying a MEL-risk, *we must both specify the mean deviation, and specify with respect to what level.*

Interpretation. An MEL-risk of L means that the average loss with respect to the value for I in a time period where the hazard does not occur is exactly L. However, there are two extreme possibilities with regard to what is to be considered as I for a hazard-free time period.

(a) Natural, or explicit, hazard-free case There is actually a naturally occurring best-case hazard-free throughput capacity $I_m + G_b$ that cannot be exceeded for the value of R, and which will occur in a time period when all goes well and no hazard occurs, and where such time periods are certain to occur. Thus in a distribution of n gains and losses about the mean per time period, in at least one of the n time periods there will occur a gain deviation G_b (with respect to the mean) up to the hazard-free case will occur, but no gain deviation exceeding G_b will ever occur. Note that the hazard free throughput capacity level implies that no variation in capacity can occur above that level, and that variation

can occur at any level below it, thus ensuring that *all possible variation is included in, and certain loss is excluded from, the MEL-risk measure*

(b) *Artificial or implicit hazard-free case* The values in each time period fluctuate about the mean I_m, and distribution of the per-period deviations from the mean follows some reasonably bell-shaped distribution, where large but usually improbable gain deviations from mean I_m do sometimes occur, and where no explicit hazard-free throughput capacity can be determined. In such a case, we may define an *artificial hazard-free case* for throughput capacity $I_m + G_b$, by defining an imaginary hazard-free time period where the gain G_b is 2 standard deviations up from the mean. We then define the MEL-risk as the mean expected loss with respect to $I_m + G_b$ for this imaginary hazard-free time period, with the throughput capacity in each time period being considered as exhibiting a loss with respect to the hazard-free $I_m + G_b$, except for the rare time period with a throughput capacity value lying beyond 2 standard deviations above the mean, which is taken as a negative loss (a gain) with respect to the hazard-free I.

In both cases MEL-risk can therefore be quite simply viewed as the hazard-free deviation, either natural or artificial, up from the mean, but also equal to the average loss to be expected in the future with respect to, or down from, throughput capacity I for the hazard-free time period (real or artificial).

When there is no natural hazard-free case, and the deviations from the mean follow a bell shaped distribution, a common situation in finance and many physical systems, the SD-risk measure and the MEL-risk measure are equivalent, since MEL-risk is exactly twice SD-risk or equal to 2-SD-risk, although they each are with respect to different standard levels.

If there is an actual hazard-free throughput capacity, with only deviations down from the hazard-free I, the distribution of deviations about the mean will tend to be skewed on the left (or truncated on the right), since upward fluctuations are blocked by the hazard-free I that cannot be exceeded, and yet very large, if rare, downward deviations from the mean of I can occur. In such a situation there seems to be no simple equivalence between the SD and MEL-risk measures. But the MEL-risk measure has an obvious advantage here, since it is precisely equal to the average loss that can be expected with respect to the hazard-free or best-case situation, and since the SD-risk measure would now be applied to a skewed distribution, something for which it is really not designed. This situation can be expected to occur frequently in systems in general, but rarely in financial systems involving stocks and bonds (but it does occur in insurance related systems); hence the need for the MEL-risk measure.

Notice that where there is exposure to future loss, where the future loss includes a certain loss and a possible loss, then the MEL-risk measure also selects out only the possible loss, that is, the true risk. Suppose again a system with ideally I = 400 if no future loss exposure, but which is actually exposed to a future loss in I whose mean is 100 and whose standard deviation is 14 where the least loss is always greater than about 70. Once more this means a certain loss of 70, plus a loss whose mean is 30 that can be as small as about 0 and as large as about 60, with a standard deviation of 14. Applying the MEL-risk measure, we have an artificial hazard-free throughput capacity 2 standard deviations up from the mean, that is at I = 328. Thus there is a certain loss of 70, or an almost certain loss of 72, plus a mean loss of 28, the MEL-risk, down from the hazard-free level of 328, giving a mean throughput capacity I_m of 328 – 28 = 300. Occasionally a fluctuation up from the mean of 300 may reach I = 330, just slightly above the supposedly (but artificial) hazard-free best-case of 328, but this may be simply considered as a "loss" of –2.

If, on the other hand, there is only exposure to possible loss, and never to certain loss, there is no need for an artificial hazard-free throughput capacity. In the example above, if for I =

400 with no loss exposure, there is exposure only to a possible loss whose mean is 30, this means the average loss down from the true hazard free I of 400 is 30, the MEL-risk, for a mean I_m = 370.

The true advantage of the MEL-risk measure is that it continually forces the user to think in terms of the distinction between certain loss and possible loss (so arguably it *is* merely an intellectual crutch), whereas the SD-risk does not. As a result, although SD-risk is absolutely correct as a risk measure, in the author's opinion, it can lead to both confusion in thinking about risk and to obscuring some of fundamental attributes of risk with aspects of systems in general described by equations (1C), (1D) and (1E).

System environments, non-limiting resources and risk

We saw earlier that if resources R of a system can be divided into R_1 and R_2 for subsystems in series, then R_2 can be operating at full capacity and R_1 not at full capacity, so that R_1 is the non-limiting resource (for the non-limiting subsystem). We saw also that if R_1 is non limiting then throughput capacity I obeys $I = K_1R_2$ for increases in R_2 up to the point where R_1 has started to operate at full capacity. However, if R_1 is so far away from full capacity that the normal range of changes upwards in R_2 will never cause R_1 to operate at full capacity, then we can neglect R_1 in figuring the effect of changes in system resources on throughput capacity I. In such a case *we can regard very non-limiting resources R_1 as part of the system environment.*

Every system operates somewhere, and it is always possible to consider the whole Earth as comprising the system. However, most of the Earth, often even the building housing, or geographic area containing, the system, will behave as a non-limiting serial subsystem, whose resources can consequently be neglected from the system. Thus *we can define the system environment as a collection resources that are non-limiting, very far from full capacity, but necessary for operation of the system.* Thus what is considered part of the system and what is part of the environment will be somewhat arbitrary. As an example, consider a computer system doing infrequent information retrieval from a small <1.0 Mbyte file on a 3 Gigabyte hard disk. Because the disk is so far from being used at full capacity it would be legitimate to regard it as part of the system environment, although most computer systems specialists would probably include it as part of the system. As another example, if a railroad track is very short, is doubled, and is currently being used infrequently by only a few trains each with exclusive control of the track, then the track is so far from being used at full capacity that it would be legitimate to consider it as part of the environment.

For a given system, risk depends on the system relative to the environment, which we denote by r(E) in this paper. If we move the system from one environment to another, the risk many change, or if we keep it in the same environment but alter the layout of the system without changing each subsystem's functionality, or if we change the subsystem functionality, risk may also change. Thus if we have different risks $r(E_1)$ and $r(E_2)$ we may either have two different environments E_1 and E_2, or we may have the same environment but a different environment relative to the system. In the following discussion *an environment E_n always means a specific environment relative to the system.* Also, equations (1B), (1C), (1D), and (1E) are valid only for systems operating in *efficient environments*, a concept to be defined later.

The influence of risk on throughput capacity

We now derive the basic risk equation (1B), and show that, for a risk measure r(E), the risk of a system in an efficient environment E relative to the system:

$$I = R[K + (b_{pb} - 1)r(E)] = RK + Rb_{pb}r(E) - Rr(E) \quad (1Ba)$$
$$= R(K + b_p r(E))$$

where the independent variables are R and r(E), and where the risk is positive, that is, it is risk it can pay to take (on average), and where, it is crucial to realize, risk measure r(E) is a measure of possible throughput capacity loss, per time period, per (basic harmonic resource) unit R, that is, *measured as a fraction of R, and not as a fraction of throughput I.* The risk r(E) is SD-risk or MEL-risk per unit R.

We also show that for negative risk:

$$I = R[K + (b_{nb} - 1)r(E)] = RK + Rb_{nb}r(E) - Rr(E) \quad (1Bb)$$
$$= R(K - b_n r(E))$$

Expression (1Ba) is a statement of the following rule: throughput capacity for a given system resource level R increases linearly with R (for valid R changes, as specified in Section 1) and also *linearly* with increases in the risk measure r(E) of distinct system environments relative to the system, provided the risk is positive and the environments are efficient. If the risk is negative, throughput capacity will decrease linearly with increasing r(E) (expression (1Bb)). If the environments are not efficient, the equation's constants b_{pb} , b_p, b_{nb}, and b_n will no longer be constant.

Efficient environments and linear relationship between throughput and risk

Now consider now two environments E_1 and E_2 relative to the system. Suppose E_1 is a risk-free environment in which the system has an unvarying throughput capacity I = KR per time unit for a system with resources R, in accordance with expression (1A) with no sharing procedure. E_2 is the same as E_1 except that in E_2 the system is in a positive risk environment.

Suppose gross throughput capacity in E_2 is KR + G per time unit, in each of one (or more) time periods in which the risk in E_2 is run but where it just happens (by good luck) that the hazard does not occur. Thus KR + G can be viewed as the hazard-free throughput capacity in the presence of risk but where the hazard does not occur. However, when a risk is run repeatedly, throughput capacity losses must occur over time. If the average throughput capacity loss per time unit, due to the hazard occurring, is L_r, then the net throughput capacity from running the risk in reality will on average be KR + G - L_r per time unit.

But expression KR + G - L_r must also give the expected or average throughput capacity I, so that we can take L_r as the MEL-risk with respect to the hazard-free capacity of I = KR + G. In general there are now two possibilities for this risk: it can be r*isk which it can pay to run repeatedly (positive risk)*, where G > L_r, or it can be r*isk it cannot pay to run repeatedly (negative risk)*, where G < L_r.

Assume now for E_2 that G-L_r is positive so that E_2 exposes the system to a risk it can pay to run repeatedly. Now recall that environment E_2 was risk free, with I = KR at full capacity for resources R applied. It is clear that there will be an increase in throughput capacity by shifting R from an environment E_1 with no risk and throughput capacity KR, to environment E_2 with average throughput capacity KR + G - L_r, that differs only in E_2 having a risk it can pay to run repeatedly.

But out of E_1 and E_2 we can construct an arbitrary number of synthetic environments, each with risk it can pay to run repeatedly, intermediate between the zero risk in E_1 and the MEL-risk L_r in E_2. We can do this in actual practice by taking a (valid) fraction of R and applying it to

E_1 and the remainder of R to E_2. For example, suppose such a synthetic environment E_S when the fraction is 50%. In E_S, for the resources R/2 operating risk free, the throughput capacity I will be KR/2, and for the remaining resources in risky operations, it will be $(KR + G - L_r)/2$ for a total of $KR + G/2 - L_r/2$, on average. Thus in E_S the MEL-risk, with resources R applied, will be $L_r/2$, and the increased throughput capacity in excess of KR, due to running the risk, will be $(G - L_r)/2$, on average. We can repeat this with any fraction, so that it is clear that for such synthetic environments, expected throughput capacity I will increase linearly with MEL-risk.

We can even have a synthetic environment E_S where the average throughput loss, and thus MEL-risk, exceeds L_r, if we include the case where additional resources are borrowed and applied to E_2; for example, if we borrow an additional R resources (at the cost KR of the risk-free throughput from the borrowed R), and apply them to E_2, the throughput capacity I is now $(2KR + 2(G - L_r) - KR)$ for the system, or $F + 2G - 2L_r$, on average, so that the extra throughput capacity for agent's resources R is $2(G - L_r)$, on average, with the MEL-risk being $2L_r$, consistent with throughput capacity increasing linearly with risk. Thus in general, average or expected throughput capacity I is given by:

$$I = KR + nG - nL_r$$

where nL_r is the MEL-risk, R and n are independent agent-controlled variables, and $n >= 0$ and may vary with the synthetic environment chosen for the system by the agent.

Now, the above expression should be pondered over, for it is absolutely correct, regardless of the distribution of losses over time, provided only there exists, among all the hazard-occurring loss-generating time periods, a time period where no hazard occurs with extra throughput capacity nG, and nL_r is the average loss over all time periods with respect to the throughput capacity in the best-case hazard-free time period.

Now suppose that for a given system, out of the set of all non-synthetic natural occurring environments in which the agent is free to operate the system, that is, accessible environments, we select an environment for which G/L_r is the highest, and let us call that environment the *reference environment* E_e. We call G/L_r the *risk efficiency coefficient* for the reference environment. Using that environment and the risk free environment we can now construct any number of accessible synthetic environments for which the gross extra gain is nG for a loss nL_r, where n is > 0, so that in every one of these synthetic environments the gross extra throughput capacity per unit of average loss, or risk efficiency coefficient, is the highest and the same as in the reference environment. Thus there will exist a set of environments, made up of synthetic environments *and natural environments*, in which the system could operate, and in each of which, for that system, the risk efficiency coefficient is the same as that of the reference environment E_e. We call this set of environments the *efficient set* based on a specific E_e, and call each of its members an *efficient environment.* An efficient environment can thus be either natural or synthetic.

From this it follows that average throughput capacity I must increase linearly with the environment MEL-risk nL_r, for all efficient environments in which the system could run, and not all merely synthetic environments, where it is positive risk (meaning $nG > nL_r$). Accordingly :

$$I = KR + nG - nL_r$$

also holds for efficient environments. [Notice that if the agent chooses, irrationally, a range of environments in which to operate the system, some of which are inefficient, the equation will not hold; instead we will have $I = KR + mG - nL_r$ where n/m can vary from one environment to another.]

Now it is clear that if we have a valid increase in R to Rf, where f is a positive real value, for example, 2.0, it is clear that this is equivalent to adding a parallel system with resources R(f-1). Hence throughput capacity I will increase to If, and throughput capacity for the risk free environment will be KRf, and gross extra throughput capacity due to the presence of risk will be nGf, and the risk will be nfL_r. From this consideration it is clear that we may legitimately be concerned with gross extra throughput capacity per (basic harmonic resource) unit R, namely ng, where g = G/R, and MEL-risk per unit R, namely nl_r where $l_r = L_r/R$, so that we can rewrite the above expression as:

$$\begin{aligned} I &= KR + ngR - nl_rR \\ &= R(K + ng - nl_r) \\ &= R(K + ng - r(E)) \end{aligned}$$

if we define MEL-risk per unit R as $r(E) = nl_r$. Since, for positive risk, $ng > r(E)$, we can rewrite ng as $(g/l_r)r(E)$, where g/l_r is greater than 1.0. Hence, for average throughput capacity I:

$$I = R(K + (g/l_r)r(E) - r(E))$$

$$= R(K + b_{pb}\, r(E) - r(E)) \qquad (2b)$$

Hence $$I = R(K + (b_{pb} - 1)r(E)) \qquad (1Ba)$$

or $$I = R(K + b_p r(E)) \qquad (2c)$$

where b_{pb} is the gross gain per unit of risk, or *risk efficiency coefficient*, for any member of the set of efficient environments, and is a constant that measures the extra gross throughput obtained for the best case of no hazard actually occurring, that is, the constant G/L_r. In practice it can be expected the while L_r is smaller than G for risks it can pay to take, it is not much smaller, so that b_{pb} will be only a little greater than 1.0, and b_p much less than 1.

Expression (1Ba) is the basic risk equation for positive risk, where r(E) is the risk of an efficient environment relative to the system. If some of the environments in which the agent (irrationally) operates the system are not efficient, equation (1Ba) still holds, but b_{pb} will no longer hold constant as the system is shifted from one inefficient environment, and therefore one risk, to another, each with a different risk efficiency coefficient $b_{pb} = nG/nL_r$. It obviously behooves the agent to discover the set of efficient environments for the system and select from that set. Since agents can be assumed to be rational and risk averse, they are not likely to run the system in an accessible risky natural environment for a smaller gross extra throughput capacity nG than could be obtained from a synthetic environment with the same risk nL_r constructed using the reference environment E_e, that is, they are likely to run the system only in efficient environments. This principle might be called the *risk equivalence principle*. It is self-evident, we believe, because its converse makes no sense. In the literature there is no sign of any research having been done into efficient system environments, a neglect it could clearly pay to remedy. For what it is worth, the author suspects, based on anecdotal evidence, that efficient system environments are highly orderly.

[K is assumed constant over long periods of time, although the expression allows for K varying (slowly in the long run) independently of the benefit of running the risk r(E) that is, independently of net addition to mean throughput capacity per unit R as measured by $b_p r(E)$. Such independence happens in financial systems where K corresponds to the risk free (per unit) interest rate, which does change over the long run, independently of the benefits of risk taking [10]. Expression (1Ba) assumes such independence of K and b_p for systems in general. If for some system it can be shown that the benefits of risk-taking have a long-run linear variation with K, as might sometimes happen, then we could write $b_p = a_pK$, giving $I = KR(1+a_p)r(E)) = KR[1 + (a_{pb} - a_{pc})r(E)]$, which is a possible variant of the risk equation.]

The quantities $Rb_p r(E)$ in (2c) and $KR(b_{pb}-1)r(E)$ in (1Ba) are each expressions for the average net extra throughput capacity achieved by taking the risk r(E) of the environment, measured as MEL-risk per unit R. Where SD-risk is preferred, so that typically SD-risk is 0.5 times MEL-risk, one can convert from MEL-risk by inserting the corresponding SD-risk measure r(E) into the equations of risk and adjusting the constants b_{pb} and b_p.

In (1Ba) KR is the unvarying throughput capacity I for a risk-free environment, and which therefore induces no fluctuations in I due to hazards. Particularly when using MEL-risk, but ultimately also with SD-risk, $Rb_{pb}r(E)$ corresponds to the gross extra throughput capacity nG in a time period where the risk is present but the hazard does not occur, and Rr(E) corresponds to the average throughput loss nL_r over all time periods due to the hazard occurring.

[In the finance arena, SD-risk r(E) is used; the equivalent of equation (2c) is also used, but equation (1Ba) is unknown. With financial systems, using:

$$I = R(K + b_p r(E)) \qquad (2c)$$

the system resources R become the principal sum invested, and K becomes the risk-free per-unit interest rate obtainable from (risk-free) Treasury bills, so that $b_p r(E)$ is the extra per-unit return gained by exposure to risk, that is, in percentage terms if K is 10%, $b_p r(E)$ might be 3%, for a total return of 13%. If the principal R is invested in common stocks, that would put the system in one risky environment with one r(E) value, if R is invested in bonds that would put the system another risky environment with a different (smaller) r(E) value, and if R is in Treasury bills that would put the system in a risk-free environment; if we distribute R over stocks, bonds and Treasury bills that puts the system in a synthetic environment with a further r(E) value. b_p varies from decade to decade, but is of the order of 0.3; this means, crudely, that a lot of fluctuation in return has to be endured to get an small increase in mean return, i.e. to get an extra 3.0 percentage points of return (of R) on average, $b_p r(E)$ must be 0.03 so that r(E) must be 0.1, so that about 10% (of R) standard deviation fluctuations in overall return must be endured. Similarly to get an extra 6.0 percentage points on average, 20% standard deviation fluctuations must be endured, and so on [20, 21, 28]. For financial systems, the SD-risk r(E) is the standard deviation in annual return per unit time per unit R, thus, the standard deviation risk is expressed as a fraction of R, and not as a fraction of the mean or expected return itself.

If investors are never irrational, then for the same return in two different environments the risk should be the same, otherwise the environment with the lower risk would be preferred and securities price levels will adjust. If one environment, say the common stock environment S, has a fluctuation level and thus SD-risk r(S) that is q times the SD-risk r(B) of another environment, say the bond environment B, so that r(S) = qr(B), then the extra return s in S above the risk free return K must be q times the extra return b in B above K, or s = qb, and securities prices will adjust to make it so. Typically q > 1.0 since stocks are riskier than bonds. This s = qb must hold, since when principal sum R is in stocks (say the S&P500 index), it is possible to construct a synthetic environment (T) with R consisting of a principal sum qR in bonds with a sum (q-1)R borrowed at the Treasury bill rate K, for which the risk r(T) is qr(B), the same as for the stock portfolio. The stock and synthetic bond-based portfolios must now each give the same return s = qb, otherwise the environment S or T with the lower risk will be preferred and securities prices will adjust to make the returns identical. It is this principle of equivalent risks giving the same return that was used originally [20, 21, 28] to derive the relation (2c), which is probably the fundamental expression of finance. This method was not used earlier in this paper, and indeed cannot be used, to derive the basic risk equations (1C) and (2c) for systems in general, since this method is based on the assumption that markets will adjust their price levels to force all financial environments, stock, bonds, treasury bills, and combinations of these, to form a single set of efficient environments for investment. In this paper the basic risk equations were derived

for all non-growth non-evolving agent-directed systems only from considerations of the fundamental nature of risk.]

It should also be clear from a similar analysis of a set of efficient environments based on a reference environment for which the ratio nG/nL_r is maximal but less than 1, that is, containing only risks it can not pay to run, that we must also have

$$I = R[(K + (b_{nb}- 1)r(E)] \qquad (1Bb)$$
$$= R(K - b_n r(E))$$

which is the basic risk equation where r(E) is a measure of negative risk for an efficient environment, or risk it can not pay to run (that is, $b_{nb} < 1$ or b_n is positive). Thus b_{pb}, for positive risk, is greater than 1, and b_{nb}, for negative risk, is fractionally less than 1.

A numerical example of the use of the positive risk equation will give the reader a better understanding of its implications. Suppose a system with R = 1000, with K = 0.075 or 7.5%, so that in a risk free environment, I is 75 x-units per week. Suppose we place the system in an efficient environment E with a MEL-risk r(E) of 0.02 x-units per week per unit R, and this placement raises mean throughput capacity from 75 to 80 x-units per week. In that case

$$I = R[0.075 + 1.25\, r(E) - r(E)] = R[\, 0.075 + 0.25 * 0.02]$$
$$= 1000[\, 0.075 + 0.005] = 75 + 5 = 80 \text{ x-units per week}$$

The best case throughput capacity is:

$$R(0.075 + 1.25r(E)) = 1000[\, 0.075 + 0.025] = 75 + 25 = 100 \text{ x-units per week}$$

from which level there are downward throughput capacity fluctuations, with a mean of 20, to mean I = 80 on average, with a minimum downward fluctuation of zero, and a maximum probably of 40, to a minimum throughput capacity of 60. If we shift the system to another efficient environment with double the MEL-risk r(E) of 0.04 x-units per week per unit R, the mean of I will be raised to 85, and the hazard free case to I =125, for a mean fluctuation of 40 down from125 to a mean I = 85. If instead the system is placed in an environment with an SD-risk of 0.02 x-units per week per unit R, throughput capacity rises from 75 to a mean of 80 as before, but with the standard deviation of fluctuations about the mean of 80 being 20 x-units per week, instead of the maximum fluctuation about the mean of 80 being about 20 if the risk of 0.02 were MEL-risk.

Risk combining with destructive interference

Readers are cautioned about naively using the risk equation to combine parallel systems with risks from different efficient environments. Suppose a composite system C made up of two parallel identical systems in two different efficient environments, with the same risk per unit R, and the same gross extra throughput per unit R, except that the risks are not well, or are negatively, correlated. Suppose throughput capacities are I_1 and I_2 such that total average system throughput capacity for C is $I = I_1 + I_2$, where

$$I_1 = R(\, K + dr(E_1)\,) \quad \text{and} \quad I_2 = R(\, K + dr(E_2))$$

Although the mean throughput capacities are additive the risks are not, since throughput capacity fluctuations in opposite directions will cancel, and thus destructively interfere. Thus for the composite system we must have

$$I = 2RK + dRr(E_1) \sim(+)\sim dRr(E_2)$$

where the tilde+ notation indicates that risk addition to give a composite risk is based on the underlying statistics, allowing for destructive interference, where:

$$Rr(E_1) \sim(+)\sim Rr(E_2) = 2Rx + 2Rr(E)$$

and where r(E) is risk per unit R for the composite (and necessarily efficient) environment E_{12} derived from E_1 and E_2, and Rx is *a positive certain loss, with respect to the hazard free throughput capacity, due to destructive interference between the capacity fluctuations of the two constituent systems.* Hence for the composite system we must have:

$$I = 2R(K + b_p r(E))$$

for any environment E, where E_{12} can be a valid value for E, where $b_p = d(r(E_1) + r(E_2))/2r(E_{12})$, that is, a value b_p that will be larger than d.

A numerical example is instructive. Suppose the two systems obey

$$I_1 = R(0.075 + 0.25r(E_1)) = R(0.075 + (1.25 - 1)r(E_1))$$

and $I_2 = R(0.075 + 0.25r(E_2)) = R(0.075 + (1.25 - 1)r(E_2))$

Suppose R = 500, and $r(E_1) = r(E_2) = 0.02$, giving mean throughput capacity of 40, and a total risk of 10, with best case of 50, for each system, and throughput capacity of 37.5 for each system in a risk free environment, so that the extra throughput capacity of 2.5 is the benefit of running the risk in each system.

If we combine the systems, the mean throughput capacity will be 80 no matter what, but if the total risks destructively interfere, so that instead of adding to 20, suppose they add to only 2, so that when $Rr(E_1)$ and $Rr(E_2)$ are each 10, $2Rr(E_{12})$ is 2. The composite system will therefore be described by:

$$I = 2R(0.075 + 2.5r(E)) = 2R(0.075 + (3.5 - 1)r(E))$$

And for 2R = 1000, and $r(E) = r(E_{12}) = 0.002$, the best case throughput capacity is 82, the mean is 80, and the total risk is 2. [This is not what would be the case if there were no destructive interference, but instead perfectly correlated risks, allowing simple addition of the risks, so that the best case is 100, the mean is 80, the total risk is 20, and b_p remains unchanged (same as d) at 0.25.] For this combined case with the destructive interference:

$$b_p = 2.5 = d(r(E_1) + r(E_2))/2r(E_{12}) = 10d = 10*0.25$$

As a result of the above analysis, it is clear that if a given system is exposed to a future loss consisting of two distinct risks r_1 and r_2, equal in magnitude, measured as MEL-risks, the resultant future loss exposure consists of a certain loss x(c) plus a risk or possible loss r_3, where

$$r_1 \sim(+)\sim r_2 = x(c) + r_3$$

where x(c) is zero only when the correlation coefficient c for the two underlying risks is +1. The certain loss factor x(c) will increase with decreasing c, with r_3 decreasing to zero as c approaches –1, the point at which all possible loss is converted entirely to certain loss and risk is zero! This is the principle of risk reduction by means of investment diversification in finance, although the fact that it merely involves converting possible loss to certain loss is rarely discussed.

Risk as a function of system environment relative to the system

The basic risk equation states that R may be altered independently of the risk r(E) (per unit system resource) of the efficient environment relative to the system. However, it should be recalled that, with equation (1A), R can be validly altered only by decreasing or increasing R in valid (basic harmonic resource) units of the existing resource type of R, so that if we n-fold increase the value of R we n-fold replicate the system. Only with such an alteration of R will r(E) not change and expression (2A) hold. If an addition of another type of resources is made we may bifurcate the system into two systems, the original system and a new one consisting of the new type of resources, and although the environment of the two systems may be identical, the *environment relative to that new system* may well be different, so that the risk per unit R may be different for the new system. In addition the environment relative to the new system may not be efficient either; it would have to be analyzed to find out.

A simple physical example should illustrate these points where risk is involved. Suppose a trucking system with resources R that ships disk drives from A to B over a gravel road designed to take 10 ton trucks at the heaviest (so that the risk of a truck being delayed because of the wheels sinking into the highway is zero), and suppose that R consists of 4 10-ton trucks each carrying 20 disk drives from A to B per day, so that I is 80 disk drives per day, and the basic harmonic resource unit is 1 10-ton truck. If we increase R by 6 10-ton trucks the risk of the

efficient environment relative to the system remains zero, and throughput capacity I climbs by 6*20 to 200 disk drives per day. If instead we had added 2 20-ton trucks each capable of carrying 60 drives per day, throughput capacity I would also increase to 200 drives per day, except that the increase in R is not valid, and worse, the 2 20-ton trucks on a road designed to carry 10-ton trucks will be exposed to risk of delay due to getting stuck. Thus there are two systems, a risk-free system where R consists of 4 10-ton trucks, for which the risk of the efficient environment relative to the system is zero, and a system where R consists of 2 20-ton trucks, with the same efficient environment (the road), but for which the environment *relative to the system* is risky, and which may or may not be efficient. In this practical context, "not efficient" would mean that there is an alternate road for the heavy trucks with less risk for the same extra throughput capacity benefit.

Combination of risk with a resource-sharing procedure.

Sometimes a system exposed to risk in an efficient environment will also have an active resource-sharing procedure at the same time to better utilize the available resources. For this case we must combine the efficiency equation (1A) and the basic risk (1B) giving, for the case of just positive risk:

$$I = R[K + (b_{pb}- 1)r(E)](1-T_S/T)[1+ sF_T(T_S)]$$

or $$I = R(1-T_S/T)[1+ sF_T(T_S)][K + (b_{pb}- 1)r(E)]$$

indicating that system resources R, in use in the risky efficient environment before application of the sharing procedure, have effectively been increased to $R(1-T_S/T)[1+ sF_T (T_S)]$ by the use of the coordinated resource sharing procedure, independently of the prior existing risk per unit R in the efficient environment. Note however, that total risk is increased to $R(1-T_S/T)[1+ sF_T(T_S)]r(E)$, so that the absolute size of losses when a hazard does occur will increase proportionally. A simple example would be a computer system with an unshared cpu executing 20 jobs per hour but exposed to the risk of corrupt input data that can cut throughput capacity to 10. If we alter the system by increasing the level of sharing, so that capacity is 100 jobs per hour, the risk is now of having capacity cut to 50, for the same level of risk r(E) per unit R of corrupt input data.

3. The preventive-resources risk equation

To derive the preventive resources risk equation:

$$I = R(1 - aP)[K + (b_{pb} - (1-N(P)))r(E)] \quad (1Ca)$$ /* for positive risk and p = 0 */

$$= R(1 - aP)[K + (b_{pb} - (1-N(P(1-p) + pP/r(E))))r(E)]$$ /* positive risk and 0 < p < 1 */

$$= R[K + (b_{pb} - (1-N(P/r(E))))r(E)]$$ /* where a = 0 and p = 1 */

with independent variables R, P, and r(E), we continue with the thread of the analysis from the previous section, since we are still dealing with system resources and risk. In what follows it is always to be assumed that risky environments are efficient environments (if the environment is not efficient, equation (1Ca) will still hold, but b_{pb} is no longer a constant).

In some environments it will be possible to apply additional *preventive resources* P to either the system environment, or to the system, to prevent loss due to risk, without affecting nG, the effect of the preventive resources being to prevent some or all of the losses nL_r. When preventive resources are added to the environment only, P is taken as the total preventive resources added. However, when preventive resources are added to the system, in cases where preventive resources must increase with increasing R to retain their benefit, we take P as resources per (basic harmonic resources) unit R. Preventive resources P may be physical or

informational, for example indexes added to files or databases, but for the present we assume then to be physical.

As an example of adding preventive resources to the system, take a simple computer system where we add x printers per unit system resources (P) to prevent a risk of deadlock [14, 17] involving printers. If only fixed resources where added to the system, then if the system is increased n-fold, the fixed preventive resources P will not be as effective, unless they are also increased n-fold; thus with preventive resources added to a system, it is preventive resources per unit R that matters. Other examples of preventive resources being added to the system are extra gas storage tanks added to marine oil production platforms (to reduce explosion risk), or backup tape drives to processors. In each case, the level of preventive resources needed is proportional to system resources R.

As an example of adding preventive resources to the environment, consider a transportation system involving trains (on non-limiting tracks that thus need not be considered as part of R), where to reduce or eliminate risk of delays due to snowslides, and thus risk of loss of throughput I, snowsheds in the environment over the tracks could be built. Clearly, if the number of trains (R) in the system is increased, the fixed amount of P is as effective as before.

It clearly makes a difference, albeit merely a minor technical one, whether nor not preventive resources are added to the environment (P) or to the system (P per unit R). In the discussion to follow, for the sake of brevity, until near the end, we assume P is being added to the environment.

In the P-free expression for positive risk:

$$I = R[K + (b_{pb} - 1)r(E)] \qquad (1Ba)$$

means throughput loss nL_r is the same as Rr(E), as we have seen, and it is this loss factor that is being eliminated, at least partly, by P. Now suppose a function N(P), which we call the *risk-prevention effectiveness function*, with value zero when P is zero, and which increases at an ever decreasing rate to 1.0, the N(P) saturation level, as P increases. We can assume that a rational agent will add preventive resources P in order of more effective resources before less effective resources, that is, minor resources to prevent frequent losses will be added first, with very great resources to prevent very infrequent small losses, being added last. It follows that nL_r or Rr(E) must be eliminated at an ever decreasing rate as P increases, so that we must have:

$$I = R[K + (b_{pb} - 1)r(E) \quad + r(E)\, N(P))$$

Notice that P does not appear explicitly in the expression. This reflects the fact that it is the risk reduction effect of P, as reflected in N(P), that matters and not the size of P. For example, two quite different physical investments P1 and P2, once more expensive than the other, might have the same risk reduction effect on throughput capacity I, so that N(P1) = N(P2). The above expression simplifies to

$$I = R[K + (b_{pb} - (1\text{-}N(P)))r(E)]$$

However, especially when the resources P are added to the system and not the environment, P may exert a slowdown effect on the system independently of its beneficial effect on reducing risk. This slowdown will have the effect of reducing the effectiveness of resources R, which should then be replaced by R(1-f(P)), where f(P) is a climbing function of P that normally has values much less than 1.0. We would expect f(P) to climb linearly with P in most cases, that is f(P) = aP where a is a small positive constant, so that the above expression would in may cases be more correctly written as

$$I = R(1 - aP)[K + (b_{pb} - (1\text{-}N(P)))r(E)] \qquad (1Ca)$$

which is the preventive-resources risk equation for positive risk, that is, for $b_{pb} > 1$. The equation states that if there is no slowdown effect (a = 0), when P reaches a level sufficient to eliminate the risk entirely, that is, when the risk-prevention effectiveness is 100%, I will revert to the hazard-free or best-case level:

$$I = R[K + b_{pb}r(E)]$$

Note that although the slowdown constant is normally small and positive, the author has uncovered a few examples where it will actually be negative, causing a "speedup" in the system and thus an increase in throughput capacity.

Note, however, that where aP is positive, the slowdown effect means that there is a value for P at which there is a maximum value for I. This maximum must occur, since the negative effect on I of adding P, due to the slowdown effect, increases linearly with P, whereas the risk reduction benefits of adding P fall off quite rapidly with increasing P. Thus at low P levels the risk reduction benefits to I of increasing P far outweigh the slowdown-effect reductions in I. At the other extreme with large P, the very small risk reduction benefit to I of increasing P is much less than the slow-down effect reduction I. Hence, at some value for P the slowdown effects of adding an increment of P is exactly balanced by the risk reduction effect of the increment in P, at which P value I is maximized. Obviously, the as the constant a approaches 0, the maximum I will approach the hazard-free best-case level above. The value for P at which I is maximized is found by solving dI/P = 0.

[As pointed out above, I does not depend on P directly. However, the value for P used does directly and negatively impact the value V generated by the throughput I, according to:

$$V = kI - C - uP \qquad (3)$$

where u is a constant(but we need to use uRP, instead of uP, if P is preventive resources per unit R). Thus if P1 and P2, with P1 > P2, are equally effective in reducing risk, that is, could generate the same I with N(P1) = N(P2), the above expression for V states that the use of the smaller P1 is preferable, from an economic viewpoint. Note also that if too much P is used, beyond that necessary for 100% risk prevention effectiveness, and saturating N(P) at unity, V is decreased more than need be via the uP term. This "too much P" has, of course, no effect on I.]

By similar reasoning , when dealing with negative risk, we will get

$$I = R[K + (b_{nb} - (1-N(P)))r(E)]$$

which, if we allow for a slowdown effect of P, becomes:

$$I = R(1-aP)[K + (b_{nb} - (1-N(P)))r(E)] \quad (1Cb)$$

which is the preventive resources risk equation for negative risk, that is, for $b_{nb} < 1$.

Note that the function N(P) is a growth function of the general form

$$G(x) = m(1 - e^{(-x/k)})$$

and saturating at G(x) = 1. If the quantity k is very small, which can also be the case, G(x) will be effectively constant and equal to 1 for practically all positive x, except x = 0, where G(x) is zero. In practice it can be expected that N(P), although having values 0 and 1.0 for zero P and large P respectively, will approximate G(x) only on average.

In equation (1C) P is an independent variable, as is R and r(E). However, if we shift the system to a new riskier efficient environment E, the effectiveness of resources of the magnitude of P in the former environment in reducing the risk in the new environment obtaining may alter. If the higher risk is merely due to a higher frequency of the same hazard, then P will likely be just as effective, so that the effectiveness of P does not diminish with increasing risk and N(P) is unaffected by changes in r(E). But if the higher risk in the new environment is altogether due to an increase in the number of hazard circumstances or in the intensity of the prior hazard, then it is near certain that the old P will be insufficient and that the P required will increase with increasing risk, most likely linearly; in such a case N(P) is no longer unaffected by a change in r(E) and should be replaced by N(P/r(E)). But in general an effect on P intermediate between these two extremes is also possible, so that in general N(P) should be replaced by the function N(P(1-p) + pP/r(E)) where p is a coupling constant ranging from 0 to 1 that controls the level of coupling between the risk level r(E) and the effectiveness of P in reducing risk, so that

$$\text{for } p = 0,\ N(P(1-p) + pP/r(E)) = N(P)$$

$$\text{and for } p = 1,\ N(P(1-p) + pP/r(E)) = N(P/r(E))$$

The most general expression for the preventive resources risk equation is therefore:

$$I = R(1 - aP)[K + (b_{pb} - (1-N(P(1-p) + pP/r(E)))r(E)]$$

This version additionally states that a riskier environment will require more preventive resources P to eliminate the risk, so that if we increase the risk we would expect a constant level of P to be less effective, that is, N() will be unavoidably smaller. Thus the growth function G(x) for N(P) in equation (1Ca) should sometimes be taken as being implicitly G(z) = G(P(1-p) + pP/r(E)) with:

$$G(z) = (1 - e^{(-z/k)}) = (1 - e^{-(P(1-p) + pP/r(E))/k})$$
$$= (1 - e^{(-P/kr(E))}) \quad \text{for } p = 1$$
$$= (1 - e^{(-P/k)}) \quad \text{for } p = 0$$

Note that mathematically, even with this extension of G(x) to G(z), G(z), and thus N(), will continue to range between 0 and 1 as is required, with any values >=0 allowed for both P and r(E), and any value between 0 and 1 allowed for the coupling constant p; in addition it will climb with increasing P and fall with increasing r(E), the sensitivity to r(E) increasing with the coupling constant, as is also required.

The case of adding preventive resources $P_{tot} = PR$ to the system, as opposed to adding P to the system environment, is normally handled in equation 1Ca in the same way as in the case of P being added to the environment.

Note that resources P are kept separate from resources R in (1Ca) in order to have an explicit method of accounting for how risk preventing-resources P affect I independently of other resources R, and also independently of the risk factors, accounted for in (1Ca) as if P were absent. Of course, if constant risk-preventing resources P have been applied, particularly to the system, but also where applied to an environment, then P can be hidden in R, with (1Ca) converted to (1Ba), as can be shown with minor mathematical manipulation of (1Ca), or P may be ignored and equation (1Ba) used to handle any residual risk.

Note also that there are two kinds of preventive resources, regardless of whether applied to environment or to the system. One type, which we might call *hazard prevention resources* P, reduces or eliminates the risk of the hazard actually occurring, as in sufficient snowsheds to prevent a track being blocked, or a high-enough dyke to prevent a flood, or sufficient water-tight compartments to prevent a ship from sinking, or sufficient computer peripherals to prevent a deadlock from occurring.

The other type of resources is *disaster recovery resources* P that come into play only after the hazard has occurred, for example, heavy equipment on permanent standby for snow removal from tracks, firetrucks at an airport, or deadlock recovery resources with a computer system. The disaster recovery resource P, while it can move N(P) from 0 to a number < 1.0 and reduce the size of the loss, will rarely be able to bring N(P) to 1.0, so that such P is rarely capable of completely eliminating loss due to the hazard occurring. Furthermore, such resources are rarely passive, as is normally the case with hazard prevention resources, but need to be used in conjunction with an emergency procedure (see Section 3).

Combinations involving the preventive resources risk equation

Consider a composite system consisting of two parallel systems each in a different efficient environment, not necessary based on the same reference environment, and each therefore exposed to different risk, and each with different levels of preventive resources, as in

$$I = I_1 + I_2$$
$$I_1 = R_1 [K_1 + (b_{ub} - (1-N_1(P_1)))r(E_1)]$$
$$I_2 = R_2 [K_2 + (b_{vb} - (1-N_2(P_2)))r(E_2)]$$

where, to keep the focus on the issue in question, we simplify by assuming that a= 0 and p = 0, that is, no slowdown effect and P unaffected by the level of r(E). There are two possibilities,

either the risks are very well correlated with correlation coefficient close to unity, or they are not well correlated. In neither case does it make any sense to try to combine the above expressions into a single expression for I. If it is the well correlated case, it is still true that the total risk can be eliminated only if sufficient P_1 and P_2 are added independently to bring $N_1(P_2)$ and $N_2(P_2)$ independently to 1.0. If they are badly correlated, the previous statement still holds, but in addition, were we to actually combine them, as in

$$I_1 = R_1K_1 + R_2K_2 + R_1(b_{ub}- (1-N_1(P_1)))r(E_1) \sim(+)\sim R_2(b_{vb}- (1-N_2(P_2)))r(E_2)$$

it would be with an expression for reduced risk because of the destructive interference phenomenon; but eliminating this reduced combination risk would give a higher increase in capacity than would be expected, since in reality at least one higher underlying risk is being eliminated. Accordingly, in both cases, it is probably best to treat each constituent system separately.

For the case where there is but one system with both coordinated sharing and a risky environment with preventive resources applied, then we need to combine the sharing equation (1A), and the preventive resources risk equation (1Ca), as follows:

$$I = R(1-T_S/T)[1+ sF_T(T_S)][K + (b_{pb}- (1-N(P)))r(E)]$$

Other combination scenarios can also be invented, but the principles outlined above, together with equation (1Ca), are sufficient to deal with them.

Use of preventive resources P and insurance considerations

Use of hazard-prevention resources P can be regarded as a form of insurance, where payment is made for the preventive resources instead of to an insurance company. However there is a significant difference, as follows. It is a fact that in very many cases a quite small physical investment P, where N(P) is far from saturation, can result in very large reductions in losses due to risk, so that savings (in value of throughput) are very much greater than the per-period cost uP. For example, in the case of a rail transportation system, small annualized expenditures for short snowsheds in places where small snowslides occur very frequently can be much less than the savings per annum from losses. But because N(P) follows the growth function G(x), where N(P) is approaching saturation, savings in losses due to incremental increased P are usually much less than the incremental expense. For example the per period expense of building a large snowshed for protection from a large snowslide that might occur once a century will exceed the average per-period savings from losses.

Instead of using physical resources P, suppose we pay a sum C to insure the losses with an insurance company. Some reflection will show that the effect of increasing P must be quite different from the effect of increasing C, since the insurance company must charge for at least the average of the expected losses (the MEL-risk) covered by insurance (usually a significant multiple of the expected losses is charged). An analysis of insurance is beyond the scope of this paper and readers are referred to the insurance literature. However, the important guiding principle that emerges from the analysis above is that it will normally be much cheaper to use a hazard prevention resources P to "physically insure" small but frequent losses (that is, where N(P) is far from saturation), whereas it is often very expensive to use incremental P for insurance against very large very infrequent losses (that is, for N(P) close to saturation), it being cheaper to buy insurance C.

Informational hazard-prevention resources P

Hazard prevention resources can be informational in nature, for example, indexes in a system for retrieval of data from a computer file or database [3, 7], or indexes in a robot system for retrieval of widgets from a warehouse of widget types. Indexes of various kinds, and other auxiliary informational structures, are commonly used in computer systems to improve system throughput capacity. Although computer scientists do not tend to consider such use as application of hazard prevention resources to reduce risk of loss of throughput [15, 27], that is exactly what is involved, since variability of throughput capacity implies risk.

Consider a simple system involving retrievals from a file of records distributed uniformly and in primary key sequence over 100 disk cylinders. If a record is to be retrieved with some fields having certain values, in the absence of an index, a sequential search must be made, so that system throughput will vary from the best case $I_m + G_b$ where retrievals are all from cylinder 0, to the worst case $I_m - G_b$ where they are all from cylinder 99, with the mean throughput capacity I_m, reflecting the time to search half the cylinders.

Thus throughput capacity fluctuates, about a mean I_m, with a clearly defined best case and an MEL-risk of G_b. If we have a primary key index, and secondary key indexes for every field in each record, then only one cylinder access will be needed for each retrieval, the risk will be eliminated and throughput capacity will be the best case of $I_m + G_b$. The indexes together are the preventive resources P. If some rarely-used indexes are omitted, the capacity will a little less than best case, and if many indexes are eliminated it will just somewhat better than I_m, all in accordance with the growth function N(P) in

$$I = R(1- aP)[K + (b_{pb}- (1-N(P))r(E)] \qquad (1Ca)$$

falling from 1.0 where all indexes are included to 0 when no indexes are included. The slowdown effect of searching the index is accounted for by the R(1- aP) factor; the bigger P, the bigger the index, and the more time lost in searching it.

Of course, what the index is actually doing is preventing unproductive application of the system search procedure to cylinders that do not contain the target records; when no index is present, the system spends unproductive time searching such off-target cylinders. Thus, whereas a sharing procedure (equation 1A) eliminates unemployment among system resources, informational hazard-prevention resources eliminate unproductive search procedures, which must necessarily employ system resources, and thus eliminate unproductive employment of system resources. Equations (1A) and (1Ca) thus also demonstrate that the major function of information is to prevent both non use and unnecessary use of resources.

Finally, the reader may ask how the obvious need to keep indexes current in the face of updating fits in here. A system that uses the file for both updates and retrievals will be able to use the indexes equally for both types of operation. But with updates the indexes may be corrupted, and risk to subsequent operations increased, if updates are to fields that are indexed. As is well known, the solution is for updates to the indexes to be made along with an update to the file [3, 7]. But this is an example of a precautionary procedure, to which the precautionary procedure equation (equation 1Da) applies. However, this particular case is a more complex one, since equation (1Ca) also applies, so that the situation is described by the combination of equations 1Ca and 1Da.

4. The risk equation with system-supported precautionary procedures

In this section, we continue with the thread of the analysis from Section 2, and derive the risk equation involving system-supported precautionary procedures:

$$I = R(1- t/T)[K + (b_{pb}- (1-H(t)))r(E)] \qquad (1Da) \qquad /* \text{ for } p = 0 */$$
$$= R(1- t/T)[K + (b_{pb}- (1-H(t(1-p) + pt/r(E))))r(E)] \qquad /* \text{ for } p > 0 */$$

with independent variables R, t and r(E).

In some environments, it will be possible to apply time-consuming precautionary procedures, that diverts resources R from normal operation for a time t in total, to prevent the losses L_r due to a known hazard, but without affecting G. These *system-supported precautionary procedures* are used to ensure that the hazard does not occur, in cases where the nature of the hazard, and where or when it may happen, *is known in advance.* In other words, there are circumstances and time periods, known in advance, during the period T for which I is computed, where the hazard is possible, that is, where risk is present, and it is possible that there are other circumstances during T, where there is no risk at all, so that precautionary procedures are useful only for a period of time $t < T$. [We assume for now that the effect of the precautionary procedure time t used to eliminate the losses due to risk will not lessen as the risk is increases, that is, we assume that the constant factor coupling t to r(E) is zero.]

Some computer system examples of precautionary procedures are the use (in theory) of the safe-state checking Banker's Algorithm [12] in operating systems prior to granting a request for resources allocation to eliminate risk of deadlock, the use of critical-section procedures in operating systems, prior to a process accessing a shared variable with mutual exclusion [18, 25], to eliminate the risk of inconsistency between cooperating processes, and use of integrity constraint checking procedures in database systems, prior to database update, to avoid risk of database inconsistency [5, 7]. Non computer-system examples are the time-consuming procedure to take a train over a weak track section or bridge in short sections, the slow-down of transport vehicles in wartime at known hazard locations to avoid booby traps, procedures for the random zigzag course of a naval vessel to avoid the risk of torpedo attack, use of de-icing procedures with aircraft in winter conditions to avoid risk of lift loss, or just the common checking & waiting procedure prior to crossing a busy road. In all of these examples, risk is present only at specific places or times, that are known in advance.

Note that the important distinguishing feature of a system-supported precautionary procedure is that *its execution requires diversion of system resources R from normal operation for a period of time.* Later we look at combinations of precautionary procedures and preventive resources P; *emergency procedures,* used after the hazard has actually occurred, are also examples of combinations of similar risk-reduction procedures and preventive resources P.

Precautionary procedures risk equation

Diversion of the system resources R for a time t (per unit R) to execute a precautionary procedure must slow down the system, and this slowdown will be equivalent to the result of taking resources $Q = Rt/T$ from system resources R, where T is the length of the time period for which I is computed. Hence where a precautionary procedure taking time t is included on a regular basis, for the case of avoiding positive risk, expected value throughput will be

$$I = R(1 - t/T)[K + (b_{pb}- 1)r(E)] = R_t[K + (b_{pb}- 1)r(E)]$$

if the precautionary procedures are ineffective and have no effect on average loss nL_r, with respect to the hazard-free throughput capacity, due to the unpredictable occurrences of the hazard. Now in the above expression nL_r is the same as $R_t r(E)$, as we have seen, and it is this factor that will be eliminated in whole or in part if the precautionary procedure is effective in avoiding throughput capacity losses due to risk. Now suppose a *precautionary-procedure effectiveness function* H(t), with value zero when t is zero, and which increases at an ever decreasing rate to 1.0, the H(t) saturation level, as t increases. We now assume that the system agent will always apply short procedures to eliminate smaller likely losses, with progressively longer procedures

being needed for larger less likely losses, and very long procedures needed for very large highly unlikely losses. As a result, nL_r or $R_tr(E)$ must be eliminated at an ever decreasing rate as t increases and the precautionary procedures are more and more effective, so we must have

$$I = R_t[K + (b_{pb} - 1)r(E) + H(t)r(E)]$$

with R_t included to reflect the decrease of I with increasing t due to reduction in the effective R because of the system time loss caused by the precautionary procedure execution. Remembering that the units of t must be time units per unit R, this expression simplifies to

$$I = R(1 - t/T)[K + (b_{pb} - (1-H(t)))r(E)] \qquad (1Da)$$

which is the precautionary-procedure risk equation for positive risk. The equation states that when the precautionary procedures are sufficiently involved and effective, as measured by t, to eliminate the risk entirely, I will revert to the best case level $I = R[(K + b_{pb}r(E)]$, less the cost due to the time delay t taken to avoid the risk, that is, it reverts to:

$$I = R(1-t/T)[K + b_{pb}r(E)]$$

Notice that if over-generous procedures are used, consuming t beyond that necessary for eliminating losses and saturating H(t) at unity, I is decreased more than need be via the (1-t/T) term. Conversely, if niggardly procedures are used that only partly avoids losses, so that H(t) has not saturated and is still less than 1.0, some residual losses due to risk will still happen, and this will decrease I, with these throughput capacity losses being accounted for by the non-zero term:

$$R(1-t/T)(1-H(t))r(E)$$

By similar reasoning , when dealing with negative risk, we will get

$$I = KR(1-t/T)[1 + (b_{nb} - (1-H(t)))r(E)] \qquad (1Db)$$

which is the precautionary procedure risk equation for negative risk.

Note that the function H(t) is a growth function of the general form

$$G(x) = (1 - e^{(-x/k)})$$

and saturating at G(x) = 1. If k is very small, which can also be the case, G(x) will be effectively constant and equal to 1 for practically all positive x, except x = 0, where it is zero. In practice H(t), while ranging between 0 and 1.0, will likely approximate the shape of G(x) only on average.

In equation (1Da) precautionary procedure time t is an independent variable, as is R and r(E). However, if we shift the system to a new riskier efficient environment E, the effectiveness of precautionary procedure time of the magnitude of t in the former environment in reducing the risk in the new environment obtaining may alter. If the higher risk is merely due to a higher frequency of the same hazard, or to the same hazard circumstances with the same frequency but with a greater associated loss, then t will likely be just as effective, so that the effectiveness of t does not diminish with increasing risk and H(t) is unaffected by changes in r(E). But if the higher risk in the new environment is due to an increase in the number of hazard circumstances, then it is near certain that the old precautionary procedure time t will be insufficient and that the required precautionary procedure time t will increase with increasing risk, most likely linearly; in such a case H(t) is no longer unaffected by a change in r(E) and should be replaced by H(t/r(E)). But in general an effect on t intermediate between these two extremes is also possible, so that in general H(t) should be replaced by the function H(t(1-p) + pt/r(E)) where p is a coupling constant ranging from 0 to 1 that controls the level of coupling between the risk level r(E) and the effectiveness of precautionary procedure time t in reducing risk, so that

for p = 0, H(t(1-p) + pt/r(E)) = H(t)

and for p = 1, H(t(1-p) + pt/r(E)) = H(t/r(E))

The most general expression for the precautionary procedure risk equation is therefore:

$$I = R(1 - t/T)[K + (b_{pb} - (1\text{-}H(t(1\text{-}p) + pt/r(E))))r(E)]$$

This version additionally states that a riskier environment will require greater precautionary procedure time t to eliminate the risk, so that if we increase the risk we would expect a constant level of t to be less effective, that is, H() will be unavoidably smaller. Thus the growth function G(x) for H(t) in equation (1Da) should sometimes be taken as being implicitly G(z) = G(t(1-p) + pt/r(E)) with:

$$G(z) = (1 - e^{(-z/k)}) = (1 - e^{-(t(1-p) + pt/r(E))/k})$$
$$= (1 - e^{(-t/Kr(E))}) \quad \text{for } p = 1$$
$$= (1 - e^{(-t/k)}) \quad \text{for } p = 0$$

Note that mathematically, as was the case with N(P), even with this extension of G(x) to G(z), G(z), and thus H(), will continue to range between 0 and 1 as is required, with any values >=0 allowed for both t and r(E), and any value between 0 and 1 allowed for the coupling constant p; in addition it will climb with increasing t and fall with increasing r(E), the sensitivity to r(E) increasing with the coupling constant, as is also required.

Execution of a precautionary procedure may result in consumption of extra resources required to make the procedure work, at a rate R_C per unit time during operation of the procedure, so that the total resources consumed is tR_C during T. This cost will have no effect on I, but must obviously affect V directly and negatively, in accordance with

$$V = kI - C - tR_C$$

A simple example of this is the case of an aircraft transportation system in winter conditions, where a de-icing precautionary procedure is used. The procedure may involve spraying the wings with anti-freeze, which is expensive, so that tR_C is significant. The cost of the anti-freeze and the amount used will not affect I as long as enough is used, but will affect V directly and negatively. (Unfortunately, where tR_C is large, with a significant affect on V (the economic "bottom line"), human operators may be tempted to run the risk.)

Finally, it needs to be underlined that with a precautionary procedure the best case hazard free throughput capacity $I = R[K + b_{pb}r(E)]$ is rarely if ever actually reached, since there is a value for t at which there is a maximum value for I. This maximum must occur, since the negative effect on I of increasing t, due to the slowdown effect of the $R(1 - t/T)$ term, increases linearly with t, whereas the risk reduction benefits of increasing t falls off quite rapidly with increasing t. Thus at low t levels the risk reduction benefits to I of increasing t far outweigh the slowdown-effect reductions in I. At the other extreme with large precautionary procedure time t, the very small risk reduction benefit to I of increasing t is much less than the slow-down effect reduction I. Hence, at some value for t the slowdown effects of adding an increment of t is exactly balanced by the risk reduction effect of the increment in t, at which t value I is maximized. The value for t at which I is maximized is found by solving $dI/t = 0$.

Classification of precautionary procedures

In practical systems, precautionary procedures appear to be of three basic kinds. They can be *check-out procedures* where conditions at a known risk point are checked carefully, for example an interactive computer procedure that checks-out all components and meteorological conditions before aircraft takeoff; a variation is the *check-out and wait procedure,* where a safe condition is awaited while being checked for repeatedly, for example the checks while waiting before enter a busy highway; another variation is *check-out and maintain or update procedures*, for example the update of an index prior to or following update of a file. Or they can be *fractionating procedures* that slow down the rate of conversion of inputs to outputs at a known risk point, by breaking down the conversion into small fractions or batches, as in taking a truck

across a risky bridge in sections. Or they can be *risk-factor removal or avoidance procedures*, as in aircraft wing de-icing prior to take-off in winter. What they have in common is that there will be risk to the system if not carried out, they are time consuming, and during execution system resources R are diverted from normal operation; it might also be added that human operators have demonstrated an inclination to omit them. The total time t the procedure executes during time T can be taken as a measure of the complexity of the precautionary procedure.

Combinations of risk preventive resources P, and a precautionary procedure

Sometimes both precautionary procedures and preventive resources are being used together. We can distinguish two separate cases.

The first case is where the system, with total resources R, is a composite of two parallel systems with resources R1 and R2, and a precautionary procedure is used with the risk in one system and preventive resources with the other. This case is quite straightforward, and follows from the combination principles laid out in Section 2. It involves algebraic combination of expressions (1Ca) and (1Da) with allowance for destructive interference, and is left as an exercise for the reader.

The second case is where the risk r(E) is being eliminated or reduced partly by use of a precautionary procedure and partly by use of preventive resources, a practice commonly recommended for operating systems to reduce the risk of deadlock [29]. This time the combination equation needed is somewhat different, and can be shown to be:

$$I = R(1- aP)(1- t/T)[(K + (b_{pb}- (1-N(P)H(t)))r(E)]$$

This equation is valid only where the resources P and the precautionary procedure time t are intertwined in such a way both P and resources and precautionary procedure time t are always required to avert incremental losses, so that if t is not large enough to saturate H(t) then increasing P to the level where N(P) saturates cannot eliminate the risk, and vice versa for unsaturated N(P) and saturated H(t). The above expression assumes that the effectiveness of neither P nor t diminishes with increasing r(E). If this is not the case, in N(p) and H(t), P and t need to be replaced by $P(1-p_p) + p_pP/r(E)$ and $t(1-p_t) + p_tt/r(E)$ where p_p and p_t are the relevant coupling constants as explained in previous sections.

An example of this, introduced earlier under preventive resources, and to which the above equation applies, is the case of a computer system that relies on a file or database index set, or equivalent auxiliary data structures P, to increase throughput capacity by eliminating the risk of long searches. In that case the index-updating procedure is the precautionary procedure. Here, even if every attribute in the file is indexed, risk is not eliminated unless the updating procedure maintains all the indexes, but even if the updating procedure maintains all the indexes, risk will not be eliminated unless the index is complete in the first instance. There is also a well-known tradeoff here: the increase in I by reduction of risk should exceed the decrease due to the system slow-down effect of updating (precautionary procedure) time t and also the slowdown effect of taking time to search the indexes which depends on aP [3, 7].

Alternatively, continuing with this second case, if the risk is such that either enough P or enough t alone can eliminate it, and where the larger the other one the easier it is, then mean throughput capacity for this combination is:

$$I = R(1- t/T)[K + (b_{pb}- (1- K(P + td)))r(E)]$$

where d is a constant and K(P + td) is a growth function of P + td that has value 0 when both P and t are 0, and which grows at an ever decreasing rate with increasing P and t to saturate at 1.0. K(P + td) is a growth function of the type:

$$G(x + zd) = (1 - e^{-(x + zd)/k})$$

This combination equation is also the one applicable to an *emergency procedure*, used after a hazard has occurred, to significantly reduce the throughput capacity losses that would inevitably occur were it not used. The time of execution of such procedures is reduced by the availability of resources P, designed, not to prevent the hazard from occurring, but to prevent excess loss when it does. Thus an emergency procedure is a time-consuming risk reducing procedure that also typically requires excess-loss-preventive resources P. The disadvantage of relying on emergency procedures is the obvious one: it is almost never possible to find an emergency (precautionary) procedure that will bring K(P + at) to saturation at 1.0, so that there are always some throughput capacity losses. The best a good emergency procedure can do is limit the losses. [Note that if the effectiveness of P + dt in reducing risk is reduced by increasing r(E), P and t will have to be modified along the lines of the modifications given earlier for P and t in similar circumstances.]

A computer example of an emergency procedure is a procedure for rapid recovery of an operating system following deadlock. A common non-computer example would is the procedure to change a wheel of a truck in a trucking system after a tire puncture – the better the tools (P) available, the less time needed for the associated emergency procedure, and thus less loss of I, but never zero loss. To entirely prevent such losses, preventive resources P alone would be needed, in the form of solid tires, as is common on military vehicles, but which, because of the extra weight, will decrease I somewhat via a reduction in the R(1-aP) term (to say nothing of the financial cost impact on V).

Combination of the precautionary procedure and sharing equations

Sometimes a precautionary procedure is applied to reduce risk in a system whose throughput capacity has been much increased by use of a coordinated sharing procedure. In that case equations (1A) and (1Da) must be combined. A little thought will show the combination to be:

$$I = R(1-T_S/T)(1-t/(T-T_S))[1+ sF_T(T_S)][K + (b_{pb}- (1-H(t)))r(E)]$$

In some cases it may be that the factor $(1-t/(T-T_S))$ can be safely replaced by the simpler factor $(1-t/T)$.

5.0 The monitoring-procedure risk equation

In the case where a monitoring procedure mechanism is necessary for reducing or eliminating risk, it is not known in advance where or when the risk will be present, that is, where the hazard could occur. It is simply known that there will be some times or places, variable from one time period T to another, where the hazard can occur, that is, where the system is exposed to risk, and the that there will be other variable times and places where the hazard cannot occur and there is no risk. It is the function of a real time risk monitoring and detection procedure, which is a component of an environment coping procedure, to detect the times or places where the risk is present in each time period and so generate an alert that triggers a response procedure, also part of the coping procedure, to take immediate action to eliminate the risk.

The risk-monitoring expression (1Ea) states that for a system with an environment coping procedure and resources R, expected throughput capacity value I can be increased by increasing a complexity-measure parameter c in a real-time monitoring procedure component of the coping procedure, in accordance with

$$I = R(1 - t/T)[K + (b_{pb} - (1 - M(c)H(t/M(c)))r(E)] \qquad (1Ea)$$
$$= R[K + (b_{pb} - (1 - M(c))r(E)] \qquad /* \text{ if risk coping time t is very small } */$$

c is an independent variable that measures the length of, and thus the complexity of, a set of constraints concisely specified in the monitoring procedure. The level of c is chosen by the system agent. These risk-meaningful constraints are meaningful with respect to risk in the environment being monitored, so that violation of a specific constraint signals presence of a specific positive risk. For each of these constraints action is taken on violation to avert the risk detected. The action is execution of either a system supported precautionary response procedure (which diverts system resources R from normal operation), or a response procedure involving reserve resources that does not divert R from normal operation. These procedures typically divert system resources R for an average time t per unit R, the level of t being determined by the agent. Thus t is also an independent variable, along with R, c and r(E). [In the above equation we are assuming for the present that the effectiveness of t is not diminished by increasing r(E).]

An environment coping procedure thus contains two components: (1) a real-time monitoring procedure with built-in capability of detecting, through violation of a risk-meaningful constraint, presence of risk of throughput loss due to a hazard in the unfolding environment, and (2) a response component consisting of a set of procedures, usually precautionary procedures, to respond to and at least partially eliminate risk of the hazards detected.

So here, unlike the case of regular and predictable application of a system-slowing precautionary procedure, as discussed in the previous section, the agent operates the system as if no risk were present, without knowledge of when or if the risk will appear, and most of the time without any significant slowdown of the system, confident that the continuously-operating environment monitoring procedure will detect risk in time to take short-term action to avoid or reduce loss of throughput.

Elimination of risk of loss of throughput capacity by means of a monitoring procedure and response procedures is a more complex case than either of the two previous cases of risk elimination by either preventive resources P or precautionary procedures alone. Nevertheless this approach to risk elimination is very important in system practice. To a considerable extent, it is actually a combination of the two previous cases, but with the preventive (detective) resources P being informational in the form of constraint specifications, as we shall see.

To show that expression (1Ea) holds generally, the first problem is to develop and specify risk-meaningful complexity for the real-time monitoring procedure. Then, for a given mean throughput capacity we need to show how increasing the risk-meaningful complexity in the monitoring procedure can increase the rate of detection of risk in advance. Once the hazard is detected, we can direct a procedure to respond to it, and so eliminate or reduce the loss-inducing Rr(E) term from the simple relationship between risk and expected value in the basic risk equation(1Ba).

In approaching the first problem, we observe that for a monitoring procedure operating in real time, the environment may unfold with a degree of unpredictability or randomness. Thus any system operating in real time will be confronted with an unfolding environment that can be characterized by at least one incoming real-time continuous data stream of bits - ones and zeros. For example, if it is a mobile robot system, there will be a data stream of sensor data from each of the robot's sensors (e.g. audio, video and tactile). For a given data stream, data that has already arrived constitutes a history of past environment unfolding, i.e. the historical data string. We begin with analysis of the simple case of a single sensor bit-stream, and then look at the general case.

Complexity measures with a simple monitoring procedure

To gain insight into principles, consider first a simple real-time monitoring system with only one incoming data stream from a single sensor. If the sensor is appropriate for detecting hazards in the environment, risk in the environment should show up as order or regularity in the data stream. A useful idea proposed by Gell-Mann [11] is that complexity is a measure of order and that the most appropriate complexity measure for a historical bit-string is an effective complexity measure C_m that must equal, or at least be proportional to, the length of a (concise) specification of the regularities or order in the string. In its simplest form we have simply a set of bit-string specifications, each of which is a regularity that can occur many times, with pointers to the locations of these regularities in the historical bit string. However, in practice nested regularities can occur within a regularity, and further regularities can occur within the nested regularities, and so on. Thus to obtain a more concise specification, the nested regularities should be removed from each regularity and replaced by pointers to specifications of the regularities removed, and so on in a hierarchical type or hierarchical explosion for each regularities. The details of this are conventional but beyond the scope of this paper; the data structure can be most concisely expressed as a many-to-many recursive database relation, with a database explosion for each regularity [3]. The length of such a specification is the effective complexity measure C_m for the historical bit string.

Gell-Mann's effective complexity C_m does not quite suit our needs in relation to a simple monitoring procedure, however, but it is close. Gell-Mann's C_m includes all regularities within the historical bit string. But what is needed is a specification of just *those bit-pattern regularities that are meaningful in revealing the presence of risk of throughput capacity loss.*. Let us call such bit patterns *risk-meaningful bit patterns*, or *rm-bit-strings*, or *rm-regularities*, for purposes of this analysis.

Note however, the incoming bit stream must not be random, that is, the cumulative bit function should not form a random walk. For a random walk the standard deviation of the function changes over time T is kT^H, where H, the Hurst constant for the bit stream, is 0.5. If H = 0.5, no meaningful regularities are possible. Meaningful regularities are possible only for H <> 0.5, in which case the cumulative incoming bit stream forms a fractal Brownian function [22]. Fractal time series is a large subject, but, although relevant for specific systems, is beyond the scope of this paper [9, 16, 23]

In order to determine the rm-regularities in a specific environment an investigator must first analyze a very long historical bit string for that environment, so long that even quite rare rm-bit patterns show up a sufficient number of times both to enable the investigator to spot them and *establish that they are correlated with a significant risk*, and to enable the investigator to measure that risk. Hence the rm-bit-strings will be a subset of all of the bit-string regularities $b_1, b_2, \ldots b_k$ listed in the root table for the hierarchy types for the regularities in the historical bit string. If we call these rm-bit-strings $r_1, r_2, \ldots r_j \ldots r_q$, where $q <= k$, then the monitoring procedure should clearly use a complexity measure c that is a concise measure of the sum of the lengths of the specifications of each r_j that the monitoring procedure is equipped to detect and respond to. These specifications can be specified as constraints, for example:

Constraint S_j : not r_j; on violation run proc$_j$

In other words, if the most recent bit string matching r_j in length is not r_j in content, there is no risk, but there is a risk if there is a content match, since this is a violation of the constraint, and consequently, the precautionary procedure proc$_j$ will be executed.

Notice that with this complexity measure c, the specification of an rm-regularity r_j does not include the specification in the root table of the pointers to locations in the historical bit string, as is the case with Gell-Mann's complexity measure C_m, since these are not necessary for functioning of the monitoring procedure. And since in addition there are fewer regularities than

with C_m, for a given historical bit string, this complexity measure c will always be smaller than the C_m measure.

Complexity measure for a general purpose coping procedure

The above discussion assumes a single data stream from a single sensor that has been selected so as to generate data that will include all relevant risk-meaningful data for the system environment. In this simplest case, any of bit patterns r_1, r_2, ...r_j ... occurring would indicate the presence of a risk, requiring execution of one of procedures $proc_1$, $proc_2$, ... $proc_j$... to avert it Thus the monitoring procedure would merely have to check for the presence of any of the r_j bit-patterns. Such a simple system is sometimes possible. For example a single sonar sensor in a submarine will detect all relevant data about mines in its path in time for evasive action to be taken.

However, very often such a simple detection system is not possible, because the physical world is usually not simple. Very often risk will appear undetected because of complex or unusual combinations of circumstances. For example, a sensed incoming bit pattern r_n that normally does not indicate the presence of risk may correspond to a chemical that in most circumstances is quite safe (e.g. a chlorate salt for use in swimming pools) but which would cause a violent explosion in contact with another chemical with risk free bit pattern r_3 that is also normally quite safe (e.g. washing up liquid, and fatal explosions of this kind have occurred). The unavoidable fact is that in either a technologically or naturally complex environment it is often very difficult for a monitoring system or humans to detect the presence of risk.

Consequently, in the general case, incoming bit string data from each sensor has to be checked against a database containing risk meaningful data about the environment. We call this database the *risk-meaningful database* or *rm-database.* The rm-database should in principle also contain all relevant data about the specific physical environment involved, and also be updated regularly or in real time from other relevant information sources. Much of the data for the rm-database will be the result of risk identification investigations into the environment – since the very first step in managing risk is the identification of the specific risks in a given environment. An example of such an rm-database is the terrain data base used in a monitoring system in aircraft to prevent collisions with mountains.

For detection of risk, a set of constraints for the rm-database can constructed, such that on attempted insertion, or merely inspection, of the current bit string section from a sensor, one or more of these constraints will be violated if risk is present. Each constraint is specified as a predicate. Violation of a specific constraint would signal a specific type of risk and trigger a procedure to deal with it.

In the general case, the constraint predicates may be quite complex and involve many different relations from the database and not just the relations dealing with the historical bit string involved in the update:

<database constraint S_j; <constraint-predicate$_j$>; on violation accept, run proc$_j$>

Such constraints may be specified in SQL [3] or equivalent database constraint language [6].

However, for complex circumstances the constraint predicates will be complex. Indeed, when video sensor data is involved and the specific color, shape or motion of an object in the environment, or any combination of these, is what signals presence of risk, the complex constraint required will in most cases be impossible to construct as a conventional relational database constraint using SQL or equivalent, and additional video image processing code will be needed; nevertheless the resulting specification will be a constraint in the most general sense and its violation will signal that risk is present. Thus in the most general case the monitoring procedure will be equipped with a set of constraints on a risk-meaningful database or set of databases, it being possible that individual constraints will be of great specification length and consequently of great complexity.

An important example of this kind of monitoring subsystem is currently being installed on U.S domestic passenger aircraft to avert collisions with mountains. The main rm-database is a terrain database of the U.S. The sensor senses the aircraft's position using global positioning technology. The position data from the incoming GPS data stream, together with altitude, speed and direction data, will cause a database constraint violation if the aircraft is below the level of a mountain top, is headed for the mountain and is within a specified distance from it; this signals an alert and requires that evasive action be taken.

Monitoring complexity and the effectiveness of a monitoring procedure

Suppose now that we denote the specification of each of the rm-database constraints for a given environment by S_A, S_B, S_C, ...S_j.... The monitoring procedure will contain the set of conditional (or prescriptive) imperatives of the form:

If <S_A violation> then <alert-A>;

If <S_B violation>then <alert-B>;

...

where Alert-A, Alert-B, ... Alert-J ... are each sets of precautionary procedures that can completely eliminate the risk detected by the corresponding rm-constraint violation. On average, although obviously not every time, the executions of a precautionary procedure set Alert-J due to constraint S_J violations will give rise to a throughput loss avoidance L_J in I.

The monitoring procedure may be said to be saturated with rm-constraints if all constraints A, B, ..., J, ..., relevant to the environment being monitored, are coded for in the monitoring procedure. If the monitoring procedure for a given environment is unsaturated, that is, not all rm-constraints for that environment are coded for in the monitoring procedure, we define the sum of the lengths of S_A, S_B,... S_J, ... as the effective rm-complexity c of the monitoring procedure with respect to that environment. Adding additional rm-constraints for that environment to the monitoring procedure will then increase the procedure's effective monitoring complexity, up to its maximum of C_S, the rm-complexity for all historical bit streams (from multiple sensors) sensed from that environment.

To derive the relationship between throughput capacity I and rm-complexity, we note that each rm-constraint S_J violation by the unfolding environment will have a specific frequency of occurrence, as evidenced by the historical data; some rm-constraint violations will occur very often, others occur only rarely. Consequently, rm-constraint S_J violations of one type may give rise to a large loss avoidance L_J on average per period T, for all executions of Alert-J, while rm-constraint violations S_K of another type may collectively give rise to only a small loss avoidance on average per period T.

Note that in this analysis we look backwards over the historical data over many time periods each of length T. An implicit assumption here is that the environment's risk statistics are reasonably stationary [4], that is, future statistics will be like past statistics, so that the expected behavior in the next future T period is the average for the previous historical set of time periods each of length T. To help the reader follow the analysis below, consider any historical time period n. In that time period there were a_{nv} violations of constraint S_A, a_{nh} of which signaled hazards that occurred, with average loss per hazard h_{na}. Also there were b_{nv} violations of constraint S_B,b_{nh} of which signaled hazards that occurred, with average loss per hazard h_{nb}, and so on. Thus, over a large number j of time periods, the average number of violations (and thus alerts signaled) due to constraint S_A is $a_v = (a_{1v} + a_{2v} + a_{3v} \ldots + a_{jv})/j$, and the average number of actual occurring hazards signaled by violation of S_A is $a_h = (a_{1h} + a_{2h} + a_{3h} \ldots + a_{jh})/j$, and total average loss due to

S_A is $L_A = (a_{1h}h_{1a} + a_{2h}h_{2a} + a_{3h}h_{3a} \ldots + a_{jh}\, h_{ja})/j$. Similarly the average number of violations (and thus alerts signaled) due to constraint S_B is $b_v = (b_{1v} + b_{2v} + b_{3v} \ldots + b_{jv})/j$, and the average number of occurring hazards signaled by violation of S_B is $b_h = (b_{1h} + b_{2h} + b_{3h} \ldots + b_{jh})/j$, and the average loss due to S_B is $L_B = (b_{1h}h_{1b} + b_{2h}h_{2b} + b_{3h}h_{3b} \ldots + a_{bh}\, h_{jb})/j$, and so on. Hence the average loss, and thus MEL-risk, for the j time periods, due to the losses signaled by violations of all constraints, is $L = L_A + L_B + \ldots L_J \ldots$, and the average number of violations signaled is $v = a_v + b_v \ldots + j_v \ldots$, and the average number of occurring hazards signaled is $h = a_h + b_h \ldots + j_h \ldots$ Furthermore the average loss for the j time periods due to the actual hazard losses signaled by violations of constraints S_A, S_B ... up to constraint S_K is $L(K) = L_A + L_B + \ldots L_K$, and the average number of violations signaled is $v(K) = a_v + b_v \ldots + k_v$, and the average number of actual occurring hazards signaled is $h(K) = a_h + b_h \ldots + k_h$. The reader may gain a clearer understanding of how these quantities relate to each other, by arranging the historical data in a diagram with time periods from left to right and losses vertical.

At this point we can usefully define a *rm-monitoring-complexity efficiency coefficient* e_m. There will be an e_m value associated with each of the constraints S_A, S_B,...S_J ... specified in the monitoring procedure. Using the historical data, for any rm-constraint S_J, e_m is the contribution to I in terms of average loss L_J eliminated (per time period T) by detection of all violations of S_J (and subsequent executions of Alert-J), divided by the length of the S_J specification. Thus

$$e_m = L_J / \text{len}(S_J)$$

This means that for a given environment unfolding over a given period of time, because of the historical statistics, we can order the rm-constraints S_A, S_B, ... in order of their decreasing monitoring-complexity efficiency coefficients. Thus the rm-constraint that over the historical period has given rise to the largest contribution to I on average per period T for the least length of rm-constraint specification or rm-complexity appears first (say type A), by virtue of its relative simplicity and high average loss avoidance L_A; the one with next largest ratio of contribution to I (say type B) to length of constraint specification appears next, and that with the lowest ratio appears last. [It should also be clear that if the order S_A, S_B ... S_K is an order that accords with decreasing rm-monitoring-complexity efficiency coefficient order, then L(K), v(K), h(K) defined above are discrete closely-correlated increasing functions, to a very good approximation, since as we go from L(K) to L(L), we would expect about the same percentage increase in v(K) and h(K) as in L(K).]

We can now define a function M(c) that is a measure of cumulative contribution to I versus c, provided we can assume that rm-complexity is added to the monitoring procedure in order of decreasing e_m. We call M(c) the *monitoring effectiveness function or schedule*. M(c) has a value between 0 and 1.0, and measures the average throughput capacity loss averted by use of rm-constraint specifications in the monitoring procedure totaling c, as a fraction of the throughput capacity loss averted when all rm-constraints for the environment have been included in the monitoring procedure, allowing all possible hazards to be detected and dealt with, that is, for $c = C_S$. More simply, M(c) is the fraction of the average loss Rr(E) due to risk that is averted by means of the rm-complexity c. Hence the losses that are averted by the level of c are RM(c)r(E). Hence if we have rm-constraint specifications totaling c, the normal average loss of Rr(E) due to risk will be reduced to Rr(E)[1 – M(c))]. Hence, on average

$$I = R[K + (b_{pb} - (1 - M(c)))r(E)]$$

where M(c) is zero when c is zero, and where M(c) climbs at a decreasing rate as c increases, until finally as M(c) approaches 1.0, large increases in c have little effect on M(c).

Note that if a specific total of rm-constraint specifications $c = c_K$ corresponds to constraint specifications S_A, S_B ... up to S_K, then $L(K) = L(c_K)$, $v(K) = v(c_K)$ and $h(K) = h(c_K)$. Thus L(c) becomes total losses on average actually signaled by constraints totaling c, v(c) is total alerts signaled for constraints totaling c, and h(c) is actual number of occurring hazards signaled by constraints totaling c.

We have so far tacitly assumed that the precautionary procedures take sufficient time to eliminate all the risks detected by the level of c. We have also ignored the cost to throughput capacity associated with time taken for precautionary response procedures. We now remedy these flaws.

During operation of a precautionary procedure, system resources R are diverted from normal operation. During time period T, the average time t per unit R taken by all the precautionary procedures to deal with the risks signaled clearly must increase in proportion to the average number of alerts generated by the monitoring system per time period T – the more alerts the more time needed to get around the risks detected, that is, t/v(K) or t/v(c) is constant. But since v(K) correlates closely with L(K), it follows that L(c) and v(c) are in constant proportion, so that t and L(c) are in constant proportion. Since RM(c)r(E) = L(c) it follows that the time t to deal fully with all alerts for the level of c must increase in proportion to M(c), that is, as c increases, t/M(c) must stay constant to a very good approximation.

In spite of this, the precautionary procedure time t per unit R expended does not depend only on M(c) since t is ultimately an independent variable controlled by the system agent (the precautionary procedure time t to completely eliminate the risks detected by the level of c depends only on M(c)). For a given level of risk detection, and thus of c, the agent can decide to expend enough precautionary procedure time to eliminate the detected risks completely, or to eliminate each of these risks only partly. But, as shown in Section 4, the fraction of the losses due to the risks detected that can be eliminated by precautionary procedures is a growth function of t, although this time, a growth function of t/M(c), namely H(t/M(c)), since the more risks detected the greater the time t per unit R needed.

H(t/M(c)) is zero for small t and increases at an ever decreasing rate to 1.0 as t increases. Also, the level of t required to saturate H(t/M(C)) increases linearly with M(c). But, as shown earlier, the losses due to the risks detected are RM(c)r(E). Hence the losses that are eliminated by precautionary procedures taking time t are RM(c)H(t/M(c))r(E). Accordingly we must have

$$I = R_t\,[K + (b_{pb} - (1 - M(c)H(t/M(c)))r(E)]$$

where R_t is the effectively reduced R due to diversion of R to the precautionary procedures..

Since R_t must be R(1-t/T), the above expression can be rewritten as:

$$I = R(1 - t/T)[K + (b_{pb} - (1 - M(c)H(t/M(c)))r(E)] \qquad (1Ea)$$

which is the basic risk monitoring equation for positive risk. Unlike the case for the other four equations, this equation has four independent variables under the control of the agent, namely R, t, c and r(E). It might be thought that H() = M(), but in general, there are no grounds for believing this, although they will be similar.

Notice that both M(c) and H(t/M(c)) must separately saturate at 1.0 to completely eliminate the risk. This is precisely what would be expected. If c is not large enough to detect all the risks (that is, M(c) < 1.0), then even if enough precautionary procedure time t is expended to eliminate all of the risks detected (that is H(t/M(c)) = 1.0), there will still be residual uneliminated risk, and vice versa (for M(c) = 1.0 and H(t/M(c)) < 1.0).

The reader is asked to pause to appreciate the profound subtlety here that is initially hard to see and easily misunderstood, but which lies at the root of the matter and cannot be evaded or expressed more simply. If a given level of risk is detected for a given level of c with M(c) < 1.0, it will take a certain level of t to bring H(t/M(c)) to (say) 0.75 and eliminate 75% of the losses detected. But if now the level of c is increased a lot, so that twice as much risk is detected and M(c) therefore now doubled but still less than 1.0, twice as much t will now be required to bring H(t/M(c)) to 0.75 and eliminate 75% of the doubled level of risk. And, nevertheless, this is what one would expect, and is provided for by expression (1Ea).

Notice also that since t is an independent variable, the system agent is free to be overcautious in the face of detected risk, and use far more precautionary procedure time t per un it R than is necessary to eliminate losses; if far too much t is used, t may equal T, so that the system is essentially thrashing (29) with no throughput capacity at all.

If the system is using some reserve resources to eliminate the risk, this will enable the time t taken by response procedures to be negligibly small, yet sufficient to saturate H(t/M(c)), so that k in the underlying growth function G(x) is very small (see below). This means that no diversion of the system from normal operation is needed to deal with detected risks, so that t is just sufficiently above 0 to permit H(t/M(c)) = 1, so that the risk monitoring equation (1Ea) reduces to its simplest form:

$$I = R[K + (b_{pb} - (1 - M(c)))r(E)]$$

In simpler terms, this equation covers the case where any risk detected by constraint violations, with constraint level set at c, is automatically eliminated in essentially zero resource diversion time.

By similar reasoning, if we are dealing with negative risk, we have

$$I = R(1 - t/T)[K - (b_{nb} - (1 - M(c)H(t/M(c)))r(E)] \qquad (1Eb)$$

which holds for the general case.

Note that the functions M(c) and H(t/M(c)) are growth functions of the general form

$$G(x) = G(c) = (1 - e^{-c/k}) \text{ and } G(x) = G(t) = (1 - e^{-(t/kM(M(c))})$$

respectively.

Both functions saturate at G(x) = 1, where the value for k is different for H(t/M(c)) and M(c). If k is very small, which can also be the case, G(x) will be effectively constant and equal to 1 for practically all positive x, except x = 0, where it is zero. In practice both M(c) and H(t/M(c)), while ranging between 0 and 1.0, will likely approximate the shape of G(x) only on average.

In equation (1Ea), both c and t are independent variables, as are R and r(E). But, if we shift the system to a new riskier efficient environment E, the effectiveness if t in dealing with risk many be diminished. The same considerations as in the precious section for precautionary procedures applies, and so for the most general case t must be replaced by t(1-p) + pt/r(E). Hence H(t/M(c)) should be replaced by the function H(t(1-p)/M(c) + pt/M(c)r(E)) where p is a coupling constant ranging from 0 to 1 that controls the level of coupling between the risk level r(E) and the effectiveness of precautionary procedure time t in reducing risk, so that

$$\text{for } p = 0,\ H(t(1-p)/M(c) + pt/r(E)M(c)) = H(t/M(c))$$

$$\text{and for } p = 1,\ H(t(1-p) + pt/r(E)) = H(t/M(c)r(E))$$

The most general expression for the preventive resources risk equation is therefore:

$$I = R(1 - t/T)[K + (b_{pb} - (1-M(c)H(t(1-p)/M(c) + pt/M(c)r(E))))r(E)]$$

This version additionally states that a riskier environment will require greater precautionary procedure time t to eliminate the risk, so that if we increase the risk we would expect a constant level of t to be less effective, that is, H() will be unavoidably smaller. Thus the growth function G(x) for H(t/M(c)) in equation (1Da) should sometimes be taken as being implicitly G(w) = G(t(1-p)/M(c) + pt/M(c)r(E)) with:

$$G(w) = (1 - e^{(-w/k)}) = (1 - e^{-(t(1-p) + pt/r(E))/kM(c)})$$

$$= (1 - e^{(-t/kM(c)r(E))}) \qquad \text{for } p = 1$$

$$= (1 - e^{(-t/kM(c))}) \qquad \text{for } p = 0$$

Note that mathematically, as was the case with H(t), even with this extension of G(x) to G(w), G(w), and thus H(), will continue to range between 0 and 1 as is required, with any values >=0 allowed for both t/M(c) and r(E), and any value between 0 and 1 allowed for the coupling

constant p; in addition it will climb with increasing t/M(c) and fall with increasing r(E), the sensitivity to r(E) increasing with the coupling constant, as is also required.

[With regard to the function M(c), it is not likely that the effectiveness of c in reducing risk will be affected by increasing r(E), especially if care has been taken to include in the function M(c) not just constraints for a given E, but constraints for all known hazards in every environment of the efficient set. However should it be the case, however unlikely, that the effectiveness of M(c) in reducing risk is also diminished by increasing r(E), then c has to be replaced by $c(1-p_c) + p_c c/r(E)$ where p_c is a coupling constant.]

There are some final economic costs to consider. The first is the cost of executing the monitoring procedure itself, typically the per unit time costs of a sensor system feeding a processor executing the monitoring procedure. The second is the per period costs of any reserve resources for the response procedures; the third is the per period cost of any resources consumed by the response procedure. The first two costs will be constant per unit time period T. However, these costs not will not affect I; what matters for I is that the systems work. They will affect V however, and can be accounted for by deducting a further constant Z, as in: V = kI - C - Z. The third cost, if incurred, must be proportional to t. This does not affect I either, but can be accounted for in V by

$$V = kI - C - Z - wt$$

where w is a constant.

Once more, it needs to be underlined that with the use of a risk monitoring and detection systems and a precautionary procedure the best case hazard free throughput capacity $I = R[K + b_{pb}r(E)]$ will often never be reached, since there is a value for t at which there is a maximum value for I. This maximum must occur, since the negative effect on I of increasing t, due to the slowdown effect of the $R(1 - t/T)$ term, increases linearly with t, whereas the risk reduction benefits of increasing t falls off quite rapidly with increasing t. Thus at low t levels the risk reduction benefits to I of increasing t far outweigh the slowdown-effect reductions in I. At the other extreme with large precautionary procedure time t, the very small risk reduction benefit to I of increasing t is much less than the slow-down effect reduction I. Hence, at some value for t the slowdown effects of adding an increment of t is exactly balanced by the risk reduction effect of the increment in t, at which t value I is maximized. The value for t at which I is maximized is found by solving $dI/dt = 0$.

6. The Second Order Effects

Second order effects are due to the appearance of conflict in systems to which a resource sharing procedure is applied to increase throughput. The classic example is a computer operating system, where the resource-sharing procedure is the direct cause of the risk of deadlock [12,14, 17], and where cooperating processes (processes sharing resources) are the direct cause of critical-section inconsistency risk[18, 25]. And the risk of deadlock for example, where it is present, will increase with the extent of the resource sharing as measured by T_S, or equivalently, as in operating systems, the extent of multiprogramming.

The author has found no elegant mathematical method of dealing with these second order effects, which might be expected given the well-known complexity of operating systems phenomena. It would appear that the best that can be done is to add second order parameters to the risk equations.

To deal with risk generated by the operation of resource sharing procedure, the constant b_{pb} and the function r(E) need to be modified, for the case of positive risk, since they can all be affected by the time T_S of the sharing procedure. The constants can be replaced by $b_{pb} + b_{pbf}(T_S)$ where $b_{pbf}(T_S)$ is a second-order effect supplement, that is a function of T_S, and where this supplement may be zero. Similarly r(E) can be replaced with $r(E) \sim(+)\sim r_p(E, T_S)$.

7. The margin of safety concept, preventive resources, and precutionary procedures

The preventive-resources risk equation allows for the important idea of a *margin of safety* or more precisely, a *preventive-resources margin of safety*. This is clearest when p = 1, where we need to use the risk-prevention effectiveness function N(P/r(E)) in equation (1Ca). Suppose now we set P well beyond the level required to push N(P/r(E)) to 1.0, so that we have excessive P and are apparently wasting resources. But the statistics from past experience may not be stationary so that the risk could actually be greater than the agent believes it to be. In that case, should the greater-than-expected hazard actually appear, the excessive P will keep the ratio P/r(E) sufficiently large to prevent N(P/r(E) from falling below 1.0. The excess in P over and above what is considered to be adequate on the basis of past statistics thus constitutes a (resources) margin of safety. Such a margin of safety, although obviously expensive, can be used in systems where throughput loss of any kind cannot be tolerated. Note that if p = 0, a resources margin of safety can be viewed as the extra P that is sufficient to keep N(P) at 1.0, even if should turn out that p is really greater than 0 after all and that, as well, the risk is really greater than expected.

Similarly, the precautionary procedure and risk monitoring risk equation also allow for the idea of a margin of safety, or more precisely, a *precautionary procedure time margin of safety*. This is again clearest when p = 1, where we need to use the precautionary procedure effectiveness function H(t/r(E)) in equation (1Da), or H(t/r(E)M(c)) in equation (1Ea). Suppose now we set t well beyond the level required to push H() to 1.0 so that we have excessive t and are apparently wasting resources. But, once more, the statistics from past experience may not be stationary so that the risk could actually be greater than the agent believes it to be. In that case, should the greater-than-expected hazard actually appear the excessive t will keep the ratio t/r(E) sufficiently large to prevent H() from falling below 1.0. The excess in t over and above what is considered to be adequate on the basis of past statistics thus constitutes a (time) margin of safety. Such a margin of safety, although obviously expensive, can be used in systems where throughput loss of any kind cannot be tolerated. Again, if p = 0, a time margin of safety can be viewed as the extra t that is sufficient to keep H() at 1.0, even if should turn out that p is really greater than 0 after all and that, as well, the risk is really greater than expected.

8. Future Research

Since the equations are valid only for agent-directed, non-growth, non-evolving systems, an obvious area for future research is development of a set of similar equations for agent-directed systems that allow for system growth and evolution. There is every reason to believe that such a set of equations could be developed. An even more challenging project, however, would be to develop a set of equations for naturally occurring, growing and evolving systems such as biological systems. The obvious problem here is the absence of any directing agent other than nature, whose impressive organizing capabilities with regard to biological systems remain to be fully explained.

8. Concluding Remarks

Five basic equations, valid for all non-growth, non-feedback agent-directed systems, but especially for computer and information systems, have been presented and justified. These basic equations may be combined, enabling expressions for a very wide variety of system situations (1F), although this area of combination circumstances may benefit from further research.

The resource sharing equation (1A) states (a) that throughput capacity increases linearly with valid increases in resources, and (b) that for a given throughput capacity, as system resources are decreased, system operating complexity must increase and vice versa. The equation

can be used to infer Spreng's triangle. The basic risk equation (1B) states that for a given system resource level, mean throughput capacity increases linearly with the risk in an efficient environment relative to the system. The preventive-resources risk equation (1C) is similar to (1B), except that, in addition, it states how expected throughput capacity will increase further if risk preventing resources are installed in the system or its environment. The precautionary procedure risk equation (1D) is also similar to (1B), except that, in addition, it states how expected throughput capacity will increase further if a precautionary procedure package is included in the system. The risk-monitoring expression (1E) is also like (1B), except that it additionally states how expected throughput capacity will increase further if a real-time monitoring procedure is in operation as part of the system. Many obvious examples of the use of these expression are to be found in multiprogramming operating systems [13, 29] and file and database systems [1, 7].

The last four equations, all of then dealing with risk, hold only for specific classes of environments, namely efficient environments. The five equations are all new except that there is a version of the risk equation that is widely used in financial systems, and which was derived by Sharpe and others in the 1960s [] on the basis of the behavior of financial markets and their participants; however the basic risk equation presented here was derived from basic nature of risk, and applies to all agent-directed no-growth non-evolving systems in efficient environments.

In all five equations the constants and variable parameters can be reduced to numbers and measurable quantities, so that the equations are subject to experimental verification. Nevertheless, the reader who has taken care to fully grasp the five equations will no doubt see that they fits with his or her own experience of practical functioning systems, especially real-time systems. However, the major benefit of the five equations to system designer and operators, is that they promote and simplify clear thinking and accurate reasoning about complex system situations and possibilities that have hitherto been in shrouded in nebulous obscurity and complexity. The equations also appear to be valid for security systems and subsystems, including computer security [29], since a system exposed to security violation is exposed to risk, although the equations were not developed with this application in mind.

References

1. Anderson, T.E., Lazowska, E. D., and Levy, H. M., The performance implications of thread management alternatives for shared-memory multiprocessors, IEEE Trans. On Computers, 38(12), 1989, pp1631-1644.

2. Barrow, J.D., "Impossibility – The limits of science and the science of limits," Oxford University Press., Oxford, 1998, p 146.

3. Bradley, J., "File and Database Techniques", Holt, Rinehart and Winston, New York, 1981.

4. Beaumont, G. P., "Probability and Random Variables", Ellis Horwood Limited, 1986.

5. Bunemann, O. P., and Clemons, E. K., Efficiently monitoring relational databases, ACM TODS, 4(3), 1979.

6. Date, C. J., Introduction to Data Base Systems, Addison Wesley, 6th Edition, 1992.

7. Date, C.J, A contribution to the study of database integrity, In C.J. Date, "Relational Database Writings", pp 1985-1989, Addison-Wesley, Reading, Mass., 1990.

8. Elton, J. E. and Gruber, M. J., "Finance as a Dynamic Process" Prentice-Hall, Englewood Cliff, NJ, 1975.

9. Falconer, K. J., "Fractal Geometry: Mathematical Foundations and Applications", Wiley, New York, 1990.

10. Fama, E.F., Risk, return and equilibrium, some clarifying comments, Journal of Finance 23(1), 1968, 29-40.

11. Gell-Mann, M and Lloyd, S., Information measures, effective complexity and total information, Complexity 1(6), 1996, 44-62

12. Habermann, A. N., Prevention of system deadlocks, Com. of ACM, 12(7), 1969, pp373-377.

13. Hennessy, J. L., and Patterson, D. A., "Computer Architecture: A Quantitative Approach", Morg an Kaufmann Publishers, Palo Alto, California, 1990

14. Holt, R.C. Some deadlock properties of computer systems Computing Surveys, 4(5), 1972, pp 179-176

15. Hsiao, D. K., and Harary, F., A formal system for information retrieval from files, CACM 13(2), 1970.

16. Hutchinson, J., Fractals and self-similarity, Indiana University Journal of Mathematics, 30 1981, 713-747.

17. Isloor, S. S., and Marsland, T.A. The deadlock problem: An overview, Computer 13(9), 1080, pp58-78

18. Lamport, L., The mutual exclusion problem, Journal of the ACM, 33(2), 1986, pp 313-348.

19. Laowska, E.D, Zahorjan, J., Graham, G.S., and Sevcik, K. C., "Quantitative System Performance," Prentice-Hall, Englewood Cliffs, NJ, 1984.

20. Lintner, J., The valuation of risk assets and selection of risky investments in capital budgets, Review of Economics and Statistics, 47(1), 1965, 13-37.

21. Lintner, J., Security prices, risk and maximal gains from diversification, J. of Finance 20(4), 1965, 587-615, 13-37

22. Mandelbrot, B.B. and Van Ness, J. W., Fractional Brownian motion, fractional noises and applications, SIAM Review 10(4), 1968, 422-437.

23. Mandelbrot, B. B., "The fractal geometry of nature", W.H. Freeman & Co, New York, 1982.

24. Markowitz, H.M. Portfolio selection, Journal of Finance, 1(1) 1952, 77-91.

25 Raynal, M., "Algorithms for mutual exclusion," MIT Press, Cambridge, MA, 1986

26. Sauer, C. H., and Chandy, K. M., "Computer System Performance Modeling", Prentice-Hall, Englewood-Cliffs, N.J., 1981.

27. Severance , D. G., Identifier search mechanisms: A survey and generalized model, ACM Comp. Survey, 6(3), 1974.

28. Sharpe, W. F., Capital asset prices: a theory of market equilibrium under conditions of risk, Journal of Finance 19(3), 1964, pp 425-42.

29. Silberschatz, A, Peterson, J. L., "Operating Systems Concepts.", 5th Ed, Addison Wesley, Reading, Mass., 1998

30. Spreng, D.T. On time, information and energy conservation, ORAU/IEA-78-22(R), Inst. for Energy Analysis, Oak Ridge Assoc. Universities, Oak Ridge, Tennessee, 1978

31. Weinberg, A.M. On the relation between information and energy systems, National Conference of the ACM, 1980, reproduced in "Maxwell's Demon", H.S. Leff and A.F. Rex, editors, Princeton University Press, 1990, p. 116.

32. Zurek, W. Thermodynamic cost of computation, algorithmic complexity, and the information metric, Nature, Vol 342, 1989, p. 119.