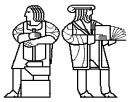

Massachusetts Institute of Technology
Engineering Systems Division

Working Paper Series

ESD-WP-2003-01.24-*ESD Internal Symposium*

LARGE SCALE INFRASTRUCTURE SYSTEMS

Fred Moavenzadeh
Massachusetts Institute of Technology

MAY 29-30, 2002

April 9, 2002

LARGE SCALE INFRASTRUCTURE SYSTEMS

Fred Moavenzadeh
Massachusetts Institute of Technology

Introduction

Highways, bridges, office buildings, houses, etc. are typical large-scale infrastructure systems of physical facilities that must be planned, designed, built, operated, and maintained. In addition to physical requirements, they have complex and often far-reaching interactions with the social, political, and economic systems they serve. Built facilities that have long service life times and large size represent a major commitment in terms of both capital expenditure and, equally importantly, social and political structures.

Large-scale infrastructure systems occur at the intersection of three systems; social systems; natural systems and technological systems. The **social system** generates demand for services; establishes the regulatory framework for their realization, operation and abandonment. It includes the economic systems that will generate the necessary resources (capital and labor) needed to build, operate, and maintain elements of the infrastructure.

Until recently, policy makers and engineers thought of **natural systems** most often as barriers that challenged human ingenuity and purpose. However, in the past few decades, it has become clear that not only is human life dependent on natural systems, but also that infrastructure planning must include measures to protect them. Large infrastructure systems must accommodate the ability of natural systems to supply the resources to build and operate them and to not exceed the capacity of natural systems to absorb the waste generated by the facilities. Recognition of "the source and the sink" capacities of nature has immense implications for the planning design, construction, and operation, and management of all kinds of infrastructure.

Technological systems enable planners and engineers to meet society's needs given the constraints of natural systems. Over the past century two major technological developments have radically altered the process of infrastructure development.

First, the mechanization of construction technology made it possible to reduce the cost of physical facilities.

The most significant elements of construction technology involved applications of first the steam engine and later internal combustion technology; from the 1920s through the 1950s, refinements of these technologies increased the productivity of construction by an order of magnitude. Reduced costs made it possible to expand highway systems, in particular the interstate highways, an important element of the automobile-based life style prevalent in the United States.

Second, information technology revolutionized the analysis, design, and operation of large-scale systems in the second half of the 20th century. The construction and technical functioning of the elements of infrastructure systems has been made far more efficient through the use of information technology. The potential of this innovation is still being realized; the management of the firms and institutions involved in construction and management of the facilities are also being rendered more effective through the applications of information technologies.

Two Types of Infrastructure Systems

Large-scale infrastructure system fall into two categories: The first consists of very large, complex, and costly systems that require all components to be in place before the system can be operational. Examples are power plants, petrochemical plants, skyscrapers, and environmental remediation facilities.

The second type is composed of many parts, each an independently functioning element requiring a small- to medium-size budget. As long as the parts are geographically dispersed, their behavior is well-understood and predictable. However, when these parts become more densely clustered geographically or interconnected operationally, they form a larger, more complex system and begin to interact with non-built systems. Infrastructure systems of this type are characterized by a network structure. Examples include transport systems (highways, railroads and waterways),

sewage systems, power distribution networks, and communication networks. Housing is another large-scale system with similar characteristics: Each housing unit is a relatively small project, but collections of them create neighborhood communities and suburbs with significant environmental, social, and political consequences far beyond those of a single unit.

Prior to World War II, most projects of the first type (commonly referred to as "megaprojects" or "macroprojects") were usually conceived by visionary individuals. The Suez Canal, Panama Canal, and Brooklyn Bridge are typical examples. The technical feasibility and cost effectiveness of these projects were often dubious. Without exception, they demonstrated substantial cost overrun, more time than anticipated to complete, and significant unanticipated political, social, economic, and environmental consequences. The effort and expense were often rationalized in terms of national pride. Fortunately, very few of these visionary schemes were ever realized.

Since World War II, some industrial projects have also reached the scope and scale of megaprojects. Working at this level, corporations building petrochemical facilities, utilities, and automobile manufacturing plants have had to look at their construction projects more holistically. The placement and construction of, for example, a new auto manufacturing plant can involve local communities, states, and even nations (e.g., Mexico and Brazil) who may use financial incentives, relaxation of regulatory systems, or favorable labor policies to entice a company to locate in its territory. At the megaproject scale, even a private sector undertaking can have complexity and far-reaching consequences for the public. Under such circumstances, the engineering and construction aspects even of a multibillion-dollar construction project are unlikely to be the decisive factors in an automobile company's planning for a new manufacturing facility.

The second type of large-scale system (e.g., those consisting of a network of many simple elements) is interesting from the standpoint of system performance. While the construction phase of such systems is not complicated, any additions to or deletions from the system often create socio-economic concerns that significantly complicate the development of the public consensus needed for their construction, operation, and

management. The NIMBY (Not In My Backyard) syndrome is a typical manifestation of public concern over any change to the status quo of these systems.

In this type of system, owners, contractors, and users usually have conflicting objectives, which require a consensus-building rather than an analytical approach to change. Since these systems are by nature very pervasive, their interaction with social and natural environments go well beyond their physical development. Transport systems are the prime example: Over the past few decades, it has not been possible to add a significant number of new transport facilities to the existing system in the U.S. The few that have been built have not been scrutinized on a conventional economic-evaluation basis--their engineering feasibility has been taken for granted. Most have been built because they meet the criterion of social desirability.

For example, on several occasions transport facilities have been abandoned or intentionally underdesigned due to social, political, or environmental concerns. Because they failed to fully appreciate the social dimensions of a project, engineers have unwittingly caused cost escalation, delay, reduction in scale and scope, and at times complete abandonment of projects. Many in the engineering profession cling to the belief that public ignorance rather than the inadequacy of the traditional engineering approach is the source of public hostility toward such projects.

Design and Construction of Megaprojects

The construction phase of megaprojects requires the concentration of large amounts of resources (manpower, equipment, and materials) in a very limited physical area over a finite number of years. This requires high level managerial as well as technological skills. In fact, the success or failure of these projects is highly dependent on the quality of the management team. Appropriate management include, in addition to having a program for the control of resource utilization, the ability to build consensus among three key parties: financier (owner), engineer (designer), and contractor (builder).

Each of these parties has a different set of objectives and interests in the project. Mostly concerned with return on investment, the financier emphasizes cost and schedule controls. However, the contractor wants profit maximization and therefore may seek change orders, which increase costs and on many occasions actually prolong completion

time. Many complex and innovative contractual arrangements and management skills have been developed to avoid the inherent conflicts among the interests of the three parties. These arrangements and skills are embedded in the socio-cultural traditions of the parties, are contextual in nature, and are not easy to transfer. Predictable and successful management of these megaprojects requires a better understanding of the dynamics of the relationships among the three parties.

An additional group whose interests must be factored in are the users of the output of these megaprojects. In industrial sectors, they are not in close contact with the facilities, which are operated by professionals.

However, in dispersed network systems, the users (e.g. drivers) are a part of the system, thus their views and reactions are a part of the system performance. In this paper, we present a strategy toward the design, construction, and management of these facilities, which incorporate socio-political considerations, users perception, and risk and reliability over the life span of the facilities.

Resource Allocation

Resources are required for realization of constructed facilities: The manner in which resources are allocated and services delivered includes several levels of detail. National- and regional-level planning must balance broad policy objectives, such as education, transportation, or public health, and directs resources toward them appropriately. At this stage, little thought is given to the individual projects these resources may support in each sector.

Per capita income, unemployment levels, literacy rate, and population growth exemplify the types of parameters national or regional policy makers may consider in determining the allocation of resources.. The increasing need to have and use such criteria is reflected in the growing usage of performance budgeting in government. However, methods to assess social and environmental factors lag sadly behind our familiar tools for economic analysis. Recent work on social and environmental accounting and the possibilities for development of “social and environmental indicators” has begun to help bridge this gap.

Once policy makers decide to direct resources toward an activity, the more detailed problems of allocating these resources to particular projects comes to the fore. For example, if government has decided to encourage regional growth and development through improved transportation, it is necessary to determine what links in the transportation chain should be established, and what mode should be employed for each link.

Decision making at this level should focus on the equilibrium of economic supply and demand. With varying degrees of sophistication, analysts can predict demand for a facility's services based upon the price a user might pay for a level of services in terms of, e.g., trip time, the cost, and the frequency of service. However, in recent years many externalities prevent straight-forward classical economic analysis. Subjective measures such as environmental, social, welfare have been introduced into the evaluation process.

The physical and operational details of the plan—such as the need for a bridge, tunnel or the characteristics of pavement and road alignment--are decided at the final stage of decision. Decision-making criteria now become such physical and technological factors as maximum vehicle load, numbers of vehicles, properties of the materials used, and anticipated weather conditions. Preference for one highway configuration over another depends on the optimization of the use of resources; decisions at this level are predicated on those at higher levels of decision making, those that determined that improved transport is desirable, and that highways are the preferred mode.

In discussing systems of constructed facilities, one may conveniently refer to the first two levels of decision as the planning stages, and to the final level as the design stage. In common usage, "planning" connotes the abstract terms in which the constructed facility is discussed; "design" addresses the "hardware" of system behavior.

Engineering practice has traditionally been concerned with design stage of decision making. Engineers have developed tools and methodologies for the analysis of systems of constructed facilities. Their job has been to provide information about the design and implementation of the activities for constructing a facility.

Planning and Design: An Interactive Process

The engineering of infrastructure systems does not concern itself with whether a particular transportation link should be a highway or a rail, or whether a particular plot within the city should have a low income or luxury housing, but rather—once these decisions are made--what physical characteristics the constructed facility must have in order to fulfill its role in the planning framework, and how these characteristics may be provided. Engineers proceed on the basis of conclusions reached by planners.

However, while the three levels of decision making have been presented as distinct stages of activity, there is, and should be, a substantial overlap between planning and design. Part of the job of decision makers in the design stage is to show when and why a system cannot meet the requirements implied in the planning decision, which itself has depended on certain assumptions. Similarly, designers may invent features for the facility that may meet the objectives of planners in a way more desirable than they had foreseen..

A major intention of this paper is to draw attention to the need for exchange of information between planning and design levels. The planners at both the national and sectoral levels must recognize that it may be impossible to implement their decisions in a manner commensurate with their own goals. In turn, the designer must be prepared to point out that sacrificing service to reduce resource requirements may not represent a satisfactory solution to the problem. Moreover, the exchange of perspectives and information should continue during the service life of the facility. Continual information feedback on how the facility is actually influencing the systems it serves may suggests changes that might be possible and desirable.

Constructed facilities can be analyzed to permit and encourage this exchange of information. This paper will explore how this analysis may be undertaken. The approach discussed may not be the only way to attack design decision problems, but is presented as one valid and useable means of doing what has not, and perhaps cannot, be done using traditional, techniques.

The approach described below is applicable to many types of constructed facilities. The examples used here refer mainly toward transportation facilities, and especially highways, but the ideas may be applied to other types of facilities.

A Concept of Performance

A system of physical facilities is intended to provide a particular set of services. The nature and volume of these services are determined in the planning process that precedes actual design. The design decisions problem is one of allocating resources to undertake actions, which will describe and bring into being a constructed facility to deliver these desired services, and to make this allocation in a most desirable manner. Questions of what is “most desirable” will also depend upon planning decisions and upon the broad role of the system of constructed facilities within the social, political, environmental and economic systems which it serves.

Design decisions are concerned with how well a particular facility provides service--its performance--and with the resources required to obtain this service--the facility’s cost. The objective at this stage is to achieve the highest possible level of performance for a given level of resource usage, and to accommodate major problems arising from the complex, multidimensional, character of both cost and performance. "Performance" may be described in terms of serviceability, reliability, and ease of maintenance.

Serviceability is a measure of how well a constructed facility provides service to the user, from the user’s point of view. "Users" are broadly defined to include not only the direct user, such as the driver of a highway vehicle, but also indirect users (e.g., merchants who receive goods shipped via highway) and subsidiary users (e.g., the eventual purchaser of the goods). Serviceability is an important indicator of the present service behavior of the constructed facility, and includes such factors as the quality of ride of a highway or the degree of comfort in the environment of a house (e.g., heat and humidity).

In evaluating serviceability, one is often dealing with users’ perceptions, and is thus dependent to a great extent upon subjective judgments. It is therefore important to continually monitor what the users want and need from a facility. Techniques for measuring serviceability are based on concepts, particularly of utility, derived from psychology and economics. A good measure of serviceability is the probability that any one of the community of users will find the services provided by the constructed facility

to be satisfactory. In practice, this parameter can be estimated by the fraction of users who judge the facility adequate.

Reliability is a measure of the probability that a facility will provide adequate service throughout the design-service life of the constructed facility. A constructed facility is an uncertain physical system serving an uncertain environment, and to make decisions based upon seemingly certain, deterministic, predictions of future service is unrealistic.

Predictions of future behavior should be made in a probabilistic fashion. Stochastic models, to be used in making such predictions, must be able to describe the way in which the facility responds to its environment, and must be able to accommodate not only the physical phenomena of weather and service usage, but also the varying effects of different operating and maintenance policies. These models may be developed through an analytical approach that includes an understanding of the processes involved when failure occurs, or through an activities approach applicable when the analyst knows only the events which lead up to an observed loss of serviceability. In either case, it may be expected that statistical data, gathered from observation of the environment and of facilities in service, will be important input to the models.

Closely related to reliability is **ease of maintenance**, proposed as a measure of the degree to which a facility will be sensitive to the uncertainties associated with future human activities. Specifically, ease of maintenance may be defined as a measure of the degree to which continued effort is required throughout the facility's service life to assure that serviceability remains at adequate levels. Maintenance activities, and the possible consequences of their neglect, represent the principal factor influencing this measure. Other factors however, such as the possible political uncertainties associated with future funding, will influence ease of maintenance and the designer's view of its importance.

Because of the primacy of maintenance in this component of performance, it is often convenient to think in terms of two distinct aspects. One is normal maintenance--the scheduled, repetitive action, which is primarily preventive in character. To the extent that normal maintenance is effective, its neglect may be expected to lead to losses of reliability and subsequently of serviceability. The other aspect is repair, referring to the

actions required if premature losses of serviceability are observed or are felt to be impending.

Together, reliability and ease of maintenance are useful measurements of the future availability of a facility. While serviceability refers to the present services provided by the facility, the other two components indicate whether these services will remain adequate throughout the remainder of the facility's design service life.At any moment in time, one may consider these three parameters–serviceability, reliability, and ease of maintenance–to determine a facility's value as a means of providing desired services throughout a specified period.

Because of the way serviceability is defined (as the fraction of users who will find the physical characteristics of service to be satisfactory) this value will ultimately refer to the chances that the constructed facility will fulfill its role in the context of social, political, and economic systems. It is at the design stage that attempts should be made to insure that a facility will exhibit the highest possible value throughout its design life, given a certain level of resources. That is, at each instant in its design service life, a facility with high value is one which exhibits adequate characteristics of present service, and the promise that these characteristics will persevere. Such a facility will be considered as acceptable solution to the design decision problem.

The performance of a facility--how well that facility provides the services for which it is intended--will be evaluated by the predicted lifetime trends of value, in terms of serviceability, reliability, and ease of maintenance. Performance is estimated by an integral, with respect to time, of value over the design life.

The Nature of Costs

Planners and designers must be concerned not only with the performance of but also with the resources required to build it, particularly the economic costs. The treatment of costs has received considerable attention in the economic literature, and an extensive discussion would be beyond the scope of this paper. A brief overview, however, is appropriate.

Basically, the cost of a resource for any particular use is determined by the most productive alternative use for that resource. For example, if concrete to be used in a

highway pavement could also be sold for use in building construction, the cost of concrete used in the pavement must be at least as high as this market price. The value of a resource is thus measured in terms of an opportunity cost, signifying the foregone opportunities for alternative uses of that resource. Moreover, as suggested above, any allocation of resources within a constructed facility may involve foregone opportunities which are not strictly economic in nature. Social, environmental, and political impacts should also be considered. Costs, like performance, are complex and multidimensional.

The assessment of alternative uses for resources is, of course, often a rather difficult task. Possible uses, and thus costs, will in general be different in the short run and in the long rung. In the long run, social and political systems can change in response to physical stimulus: Given a longer period of time, people will learn new ways of doing things. Planning decision makers must consider the relatively long service life of constructed facilities; in fact, the construction of many facilities may be justified only on their utility over a long period of time.

Another difficulty lies in the use of a common scale of measure for costs. Unfortunately (or, at least, inconveniently) not all elements of resource evaluation can be translated into monetary equivalents. Attempts to place values factors such as travel time and loss of life are generally open to serious practical as well as moral question. It seems doubtful that one measure of cost will accommodate all aspects of resource usage for constructed facilities. Despite these difficulties, it remains necessary to estimate costs. Some ways to do this are suggested below.

In the first place, resource requirements, and thus costs, for a constructed facility should be predicted for its entire design service life. The “life cycle” cost of a facility includes not only the initial implementation expenses, but also all future expenditures associated with it.

Insofar as these costs may be expressed in monetary terms, both current and future expenditures may be made commensurable by expressing them in terms of their present value. That is, future costs are discounted by opportunity cost of capital, which is usually expressed as a percentage rate. (Under certain conditions this rate will be identical with the market rate of interest on loans.) In general, such discounting is done

by referring all economic costs to a common time, usually the present, and summing to find the present value of total (economic) cost.

The problem of short versus long run is reflected in the basic dilemma of whether costs should be judged by past experience or by an estimate of future conditions. Use of past experience may be biased by conditions which the future may not replicate, thus creating the possibility that past conditions may recur. If little is expected, little is likely to be achieved. A bet on new future conditions runs the risk of being unrealistic, but can provide a stronger test of a facility's ability to fulfill the role for which it is planned. Moreover, projections based on anticipations of the future may in some degree serve to pace activity over the design service life. For example, allowing for higher maintenance expenses for a highway may encourage better quality maintenance activities in the future.

The problem of costs not readily expressible in economic terms may be partially circumvented by comparing scenarios to a single base-level alternative. Increased costs may be measured in terms of the sacrifices made in some aspects of service to achieve increases in other aspects relative to the base level values. In this way it may be possible to avoid having to define a distinct scale of measure for the costs in question. For example, the social costs of several alternative highways in terms of environmental damage, community disruption, and undesirable growth patterns, might be rated through a set of qualitative judgments comparing each option to the mean or minimum economic cost alternative.

Of course, the idea of relative cost may be applied to economic factors as well. In some cases, especially at early stages of decision making, much of the work involved in deciding upon the actual values of the proper costs of resources can be avoided with no loss of validity. Furthermore, a distinction in design must be made between costs as they are perceived by the user and the actual total cost of a facility. The point is perhaps best made through an example.

Suppose that the problem is to design a highway as a toll road, and that the resources available to be used for this system are limited to 50 percent of the revenue the road will generate. The level of toll and the anticipated number of users are determined in planning through a projection of demand. Then, suppose that comfort is the only measure of serviceability and that road roughness (as recorded with a standardized

instrument) is the only indicant required for a prediction of comfort. One might then expect serviceability to differ with varying levels of toll. At a higher toll, the user is less likely to tolerate roughness and discomfort.

Available information and experimentation will produce an estimate of serviceability versus roughness, assuming that other factors (i.e., toll) are roughly the same as the user's past experience. Recognizing this approximation, it will still be expected that at a given toll level, there is some roughness above which the predictions of numbers of users, and thus of revenues, become quite reasonable. By determining the failure point, it is possible to determine what physical characteristics of the road will be adequate.

The designer may in some cases find that the resources available are insufficient to achieve maximum serviceability. It may be necessary to devote 60 percent of toll receipts to building the road. Another policy alternative is to derive additional funding from sources external to the system in question. As long as the toll is not increased, the users' ideas of resources are unchanged, and the planning prediction holds. If the designer cannot provide the required smoothness because resources are inadequate, the planning prediction is likely to become invalid. In short, the extent to which users receive and are concerned with resource allocations–costs–is reflected in the performance measure. The design decision, however, is based upon a broader view of all of the resources required for the facility's realization.

A discussion of costs could be expanded substantially. However, the principal points to be made about costs may be quickly summarized: The costs associated with a system of constructed facilities must be include those incurred throughout the design service life of the facility, and their distribution over time can have an impact on decisions. These costs are determined by the most productive alternative uses for the resources required in realization of a facility, and as such are often complex and difficult to evaluate. The total cost of a facility cannot always be evaluated in strictly monetary terms.

Capital has a time value; to the extent that resources expenditures are expressed in terms of monetary values, they should be referred to a common time through discounting. If recognized expenditures are difficult to quantify or to evaluate in absolute terms, this

approach offers a possible solution. Comparing a facility's total cost with its performance enables the decision maker to make a rational choice among alternatives. The following section will examine the activities of this comparison at the level of design decision making.

How Decisions are Made

A system of constructed facilities is evaluated in terms of its performance and cost, i.e., the qualities of its service and the resource requirements to provide this service over its design service life. As described above, designers will make an effort to create a facility exhibiting the best performance possible at a given level of resources; as we have seen, performance and cost are complex and multi-faceted. How can decision makers compare alternative facilities be compared, and select the "most desirable"?

A design decision is made in a progressive manner, involving several analytical stages. The first stage is part of the search for an acceptable solution. Once a possible alternative with adequate qualities of service has been proposed, It is necessary to compare the resource allocations required. There will generally be several individual subscales of serviceability, as it is estimated in practice. Unless there are prestated dominance relations among these subscales, it will be most efficient to have equal estimates of serviceability on all subscales. That is, an allocation of resources which produces, for example, a high predicted fraction of satisfied users with respect to comfort on a highway and a relatively low value with respect to safety will be considered a somewhat inefficient use of these resources because the overall serviceability is likely to be limited by the lower subscale. A more efficient allocation would be made by sacrificing serviceability on the higher susbscale if this would to raise serviceability on the lower one.

It might not always be possible to balance the allocation of resources in this way, and various alternatives will be characterized by their high predicted serviceability on certain subscales. The critical point is that the balancing must be carried out to assure that at least minimum levels of serviceability are provided. This phase of the decision problem may be termed, in Herbert Simon's words "satisficing".

The next decision stage is reached when a number of alternatives have survived the satisficing strategy, each having its own cost and performance characteristics. Now, the designer will search for those alternatives that exhibit the best performance at each level of cost. If statements of cost and performance were straightforward and single-valued, this search would hardly be worth mentioning, but this is seldom the case.

The performance function will often be implicit in the design decision factors. For example, types of criteria expressed in the actual statement of performance might include, for example, minimum direct user cost at a given reliability; ease of maintenance preferred to reliability at any particular expected level of serviceability; or highest utilization of labor. By comparing alternatives with one another, within the context of such criteria, one will be able to identify best performance at a given cost, and will build up a sort of production function of increasing performance and cost.

This second screening results in a set of alternative possible solutions, all of which are acceptable and represent efficient uses of the particular resource requirement. The final selection of one facility from among this set of possibilities cannot be made within the context of the design decision problem, but must be made in view of the social, political, and economic systems with which the facility interacts, that is, at the planning level.