Proc. of SPIE Vol. 1532, Analysis of Optical Structures, ed. D C O'Shea (Dec 1991) Copyright SPIE

An Overview of the Finite Element Method in Optical Systems

Alson E. Hatheway

Alson E. Hatheway Inc.
Engineered Products and Systems

1. INTRODUCTION - A HISTORICAL PERSPECTIVE ON THE ORIGINS OF THE FINITE ELEMENT METHOD.

The finite element method is today strongly identified with structural mechanics and it was, in fact, developed in the middle third of this century by the structures community in order to achieve useful solutions in the field of the general theory of elasticity. It is often overlooked however that in the preceding centuries that were required for the development of the general theory of elasticity, the effort was driven by and often spearheaded by the very physicists and mathematicians who were hypothesizing about molecules, gravity, electro-magnetics and all the other disciplines that today make up the over-all field of "mechanics." It is not possible to appreciate the present status of the finite element method without understanding its origins.

1.1. The Seventeenth Century

The father of structural mechanics was Galileo Galilei who in 1638 posed "Galileo's Problem" (Love, 1927),

If one pushes on the end of a beam, how much will it move and when will it break?

At the time he posed the problem none of the laws or rules of mechanics that we use today in structures were in place. His conjecture however inspired three and one-half centuries of continuous effort to develop, codify and generalize the laws of solid mechanics and apply them to real structures.

Twenty-two years later Robert Hooke formulated "Hooke's Law." The law stipulates a linear relationship between force and displacement in Galileo's Problem. This was the first step in the long journey of the development of linear elastic theory. The second step occurred in 1680 when a French theoretician, E. Marotte, applied Hooke's Law to Galileo's Problem. This represents the first coupling of a theoretical principle to a practical structural problem.

1.2. The Eighteenth Century

With the first steps taken the beam theory developed quickly in the following century. James Bernoulli defined the "elastica" (what we now call the neutral axis) in 1705 and Daniel Bernoulli in 1742 suggested that its equation satisfy the condition of minimum strain energy,

$$[\mathrm{int}(\mathrm{curvature})^2\ \mathrm{dl}\ ]_{\mathrm{minimize}}.$$

Leonard Euler used this relationship to derive the equation of the elastica in 1744. In 1756 he inferred elastic instability (buckling) and in 1778 he defined the critical length of columns on this basis.

Charles Augustin de Coulomb was the first to apply "free body diagrams" to portions of a beam (1776) and using this method he determined the true position of the "neutral line" (neutral axis), determined the correct moment of elastic forces and was the first to consider the effects of shear stresses and strains. In 1787 he propounded Coulomb's Law of Torsion (applying Hooke's Law to shear stresses and strains in the torsion of beams). He also suggested the second polar moment as the appropriate section property for torsion (true only for circular sections).

The eighteenth century developments of beam theory were not capped until 1807 when Thomas Young published A Course of Lectures on Natural Philosophy and the Mechanical Arts in which he defined a tensile elastic property obeying Hooke's Law (Young's Modulus). He also applied Hooke's Law to shear stresses and strains thereby providing the definitions for the final steps in the development of a theory capable of calculating the stresses and strains in a beam and solving Galileo's Problem.

The whole process took one hundred and sixty-nine years and seemed to raise more questions than it answered. If beams can be shown to obey certain physical laws what about the behavior of two and three dimensional members?

1.3. The Nineteenth Century

In 1707 Isaac Newton had proposed a molecular concept for the micro-structure of matter. This concept allowed the development of stresses in materials based on the balance of the forces between molecules.

It was 1821 before Claude Louis Marie Henri Navier was able to extend the elastic principles of Robert Hooke to a three dimensional continuum and suggest the first complete set of equations for static equilibrium and vibration in an elastic solid. He relied on the molecular theories of Newton, Boscovitch, Laplace and Poisson and inferred the boundary conditions from the internal states of the solid, a rigorous justification of which was still nearly one hundred years in the future.

A year later, in 1822, Augustin Louis Cauchy discovered most of the elements of the pure theory of elasticity in isotropic solid bodies. Cauchy's theory used two elastic constants where Navier had used only one. The argument over the number of elastic constants in a material was to continue for several years.

In 1828 and 1829 Simeon Denis Poisson applied the general theory to numerous problems and inferred the ratio of lateral strain to longitudinal strain, later to be known as "Poisson's Ratio," but predicted the wrong value.

George Green proposed the theory of the conservation of energy and derived the elastic equations from it in 1837. His development included two elastic constants for isotropic materials, supporting Cauchy's views.

Finally, in 1845 Sir George Stokes identified compression and shearing resistances as the two elastic constants in isotropic bodies.

In 1855 Lord Kelvin (Sir William Thompson) based Green's strain energy on the first and second laws of thermodynamics. In 1862 Sir George Airy introduce the Airy Stress Function for two dimensional continua and in 1870 James Clerk Maxwell extended it to three dimensions. Maxwell also related the stress components to the strain components and related the second (spatial) differential of the strain components to a system of linear equations; conditions which are necessary to secure that strain components shall correspond to displacements (1870).

At this time the general theory of elasticity was largely complete but many previous solutions, mostly in beam theory had to be shown to be consistent with the general theory. At the same time many investigators were making new developments within the general theory. In 1855 Barre de Saint-Venant brought the torsion of beams within the general theory and the following year he did the same for the flexure of beams. In 1859 Gustov Robert Kirchhoff brought wires and spiral springs into the general theory (Kirchhoff's kinetic analog). In 1863 Saint-Venant again contributed with his demonstration of "the elastic equivalence of statically equipollent systems of loads" (Saint-Venant's Principle). In 1882 Heinrich Hertz defined and solved the contact stress problem (Hertz stresses) and John William Strutt, Third Baron of Rayleigh, developed the theory of the vibration of plates and shells. Over an extended period from 1850 to 1883 Kirchhoff worked plates and shells into the general theory.

The efforts of two hundred and fifty-four years were crowned by the publication in 1892 of A. E. H. (Alexander) Love's A Treatise on the Mathematical Theory of Elasticity, a work that had its fourth and final edition published in 1927 and is still used as the primary source in most graduate engineering schools in the study of solid mechanics and structures. Love's work documented the consolidation of structural mechanics and through its consistent notation and clear exposition provided a basis for teaching subsequent generations the general theory of elasticity in solid mechanics. His work was predominantly a work of technological synthesis: organizing and reporting on the activities of other contributors in order to show the unified nature of their work. When Love published his work mechanics was a unified and universal field for physicists. If a person worked in mechanics he was probably as comfortable with electromagnetic flux tensors as with stress tensors. In fact, the progress in one field often fueled a comparable advance in another.

It is significant that the development of the theory began and ended with two giants of optics and optical physics, Galileo Galilei and James Clerk Maxwell. A casual scan of the names of the individuals who contributed to the development and consolidation of elastic theory will disclose names that became truly famous for their contributions to optics and electromagnetic theory; Galileo, Newton, Fresnel, Thompson (Kelvin), Kirchoff, Airy, Maxwell, Hertz, Strutt (Rayleigh) and many, many more. From the historical perspective of the twentieth century it might be suggested that the discipline of structural mechanics was "created" by optical scientists that needed it to support the components in their instruments and experimental apparatuses.

1.4. The Twentieth Century

It might seem that with a complete theory in hand and with many examples solved in closed form the field of elastic solid mechanics could be considered as completely conquered. This was far from the truth. Nearly all of the practical solutions of the theory were in beams, one dimensional structural members. All the effort put into the general theory and its consolidation did little to expand beam theory beyond where it was in 1807 even though it placed beam theory on a much sounder theoretical footing. There were only a few solutions of the theory for plates and shells (two dimensional members) and virtually no solutions for solid members (three dimensional members) outside of Hertz' work in contact stresses. The general theory was couched in partial differential equations that were very difficult or impossible to solve with the complex boundary conditions of practical structural members. So, although the theory was complete the solutions were not, except for very restricted boundary condition cases such as beams. Work turned immediately to solving the boundary value problem.

In 1909 Walther Ritz published his work showing that the boundary conditions of a finite sized portion of an extended medium may be calculated from knowledge of the internal states of the medium (Bathe, 1976). His method was applied to many known shapes with assumed internal states of stress to determine the boundary conditions (loads) required to generate the internal states. Unfortunately, this was just the reverse of the way most structural problem are formulated and was little help to designers if their forms did not resemble those that had all ready been analyzed.

This difficulty was resolved in 1943 by Richard Courant who rigorously showed the inverse, namely, that the interior states of a finite portion of a continuum may be uniquely calculated from the known boundary conditions. This development inspired many workers in mechanics to develop routines using the new Ritz analysis method. By the early 1950s landmark papers were being presented on solutions in continuum mechanics, mostly in the mechanics of solids, which combined the new method with numerical integration and linear analysis techniques to calculate the displacements and stresses in structural members that were heretofore intractable. The work was labor intensive requiring many thousands of calculations using electro-mechanical calculators and perhaps an electronic analog "computer;" an appropriate time for a new computational engine to appear.

The electronic digital computer provided an engine to fulfill the promise of the new computational procedure by being able to automate the many millions of calculations previously performed by individuals with calculators. By 1960 the development of methods for solution of problems requiring the general theory of elasticity was essentially complete; even the necessary machines were beginning to be available to ease the computational effort. In that year R. W. Clough coined the new name for this method in his seminal paper "The Finite Element Method in Plain Stress Analysis." All that remained was to train people to use the new machines with the new procedure. The efforts have been extraordinarily fruitful, giving us the finite element method of structural analysis that we know today.

2. CURRENT STATE OF THE ART

In this century, following the consolidation of structural mechanics, the field was nearly abandoned by researchers interested in non-structural applications. The result has been that nearly all developments of the method have been tailored to the solution of structural mechanics problems. At the same time, the field of mechanics, which was unified as recently as 1892, has fragmented into a growing series of disciplines, each identifying its branch of mechanics with a modifier: structural mechanics, fluid mechanics, mechanics of heat, wave mechanics, quantum mechanics, etc. So that today we have an array of disciplines which are self sufficient and independent bodies of knowledge with their own cultures, lexicons, leaderships and supporting infrastructures including computational methods.

Today it seems that the advances in one discipline are usually ignored and often resisted by the other disciplines, a parochialism that is justified on the basis of the success of the current techniques. In recent years it has become common for the disciplines to cooperate in projects requiring interdisciplinary evaluations by solving the equations of their disciplines using the software of their own choice and then passing the output files to each other for interpretation and assessment. The "interpretation and assessment" may take the form of a review of the files to determine expected ranges of values (such as normal displacements) to see if they exceed levels at which they may contribute unfavorably to performance (such as OPD). More commonly, however the "interpretation and assessment" will involve additional computer processing of the output file, using it as source data for another computational procedure or code.

This approach is a natural adaptation of the "expert panel" technique for problem assessment, in which the experts gather in a conference room and present their respective opinions until a position is reached that all can agree will meet the requirements of the intended project. This approach has simply been adapted to the computer age and named "Integrated Analysis."

2.1. Integrated Analysis

In large companies one may find the Integrated Analysis method completely automated with a central piece of software which may be called a "data base manager" and is capable of running each of the analysis codes. The data base managers may also be set up with routines for converting output from one code into input for another code. With the conversion routines in place and a suitable macro-analysis file prepared it becomes possible to automatically run very complex multi-disciplinary problems through a number of separate analysis codes using the data base manager and the information in the macro-analysis file (SINDA to NASTRAN to MATRIX$_X$ to CodeV, for example). One such system is shown schematically in Figure 1 which

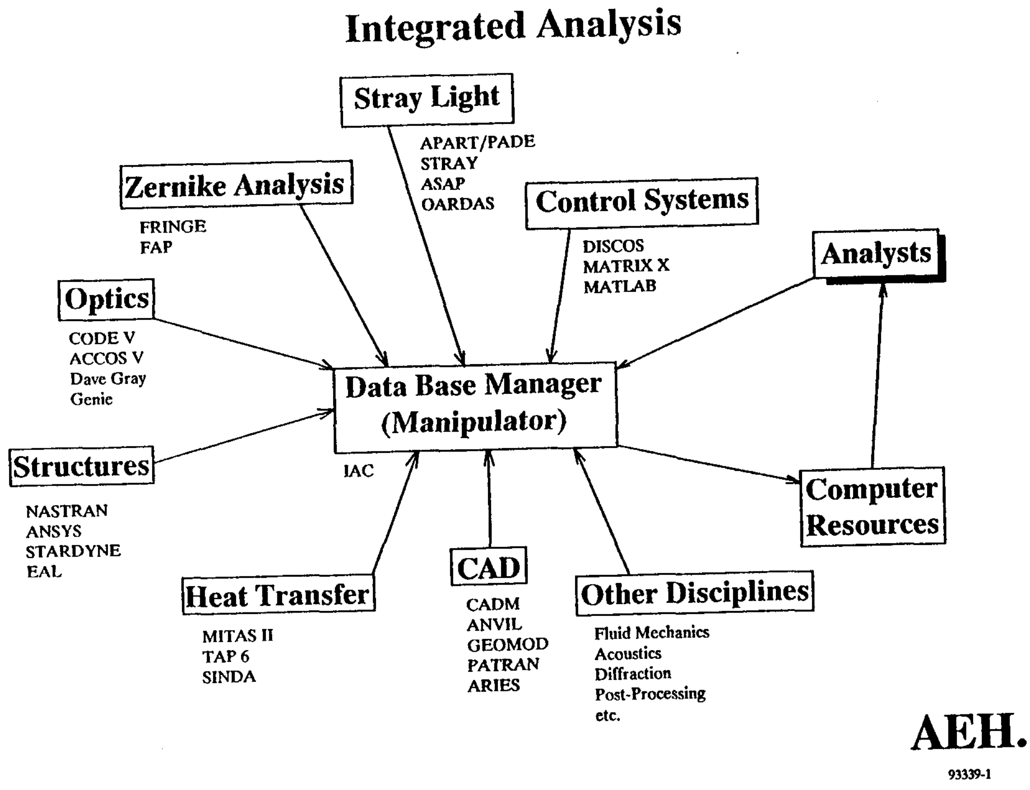

Figure 1. Integrated Analysis interfaces the software of all the disciplines to solve multi-disciplineary problems.

is modeled after a data base manager called "Integrated Analysis Capability" (IAC) which was written for NASA by Boeing Aerospace and is available through the COSMIC software distribution services.

In Integrated Analysis the analysts of each discipline are responsible for setting up the input files for the codes that calculate the performance of the system in their individual disciplines. These discipline-oriented analysts are represented in Figure 1 by the disciplines in boxes arrayed around the data base manager. A few of the various codes that might be used in each discipline are indicated also.

Additionally there is a group of "analysts" who write the conversion routines and macro-analysis files that manipulate the data among the codes. It is their job to allocate resources and keep the system running smoothly, a non-trivial task. These "analysts" are shown above the computer resources in Figure 1.

To understand each person's role in Integrated Analysis we should look at a typical analysis using this as a tool. Let's assume our problem is a large aperture space-based optical telescope in Earth orbit. Let us further assume that we are concerned about the influence of variable heating rates on the performance of the telescope. This will require a thermal analysis of the telescope under the influence of the heating from the Sun, Earth and possibly the Moon. These temperatures will be used as loads in a model of the elastic structure of the telescope and will give the displacements of points in the telescope versus time in response to the time varying temperature field. The displacements of the surfaces will be used in an optical design code to evaluate the change in optical performance caused by the time variable heating and cooling of the telescope in orbit. However, since the telescope has "active structural control" we will also need to pass information on the thermo-elastic behavior of the telescope to a control systems design code so the control systems engineers may assess the performance of their controller; this will require the subsequent passing of data from the control systems code to an optics design code for the evaluation of the change in performance.

Let us look in detail at some of the steps we have just referred to:

I Thermal Analysis (orbital transient):

Radiation, Convection, Conduction = 200 nodes
<u>Expand 200 temperatures to 1500 temperatures.</u>

II Elastic Analysis (thermo-elastic distortions):

Structural model with 1500 nodes
<u>Reduce 1500 x 6 = 9,000 DOF to 25 eigenvectors (DOF).</u>

III Control System Analysis:

25 eigenvectors plus control system functions
<u>Estimate equivalent loading condition to simulate the thermal loading.</u>
<u>Expand 25 eigenvectors to 9,000 displacements.</u>

IV Optical Analysis with 1,000 ray intercepts:

100 rays and 10 surfaces
<u>Interpolate 1,000 normal displacements.</u>
<u>Expand 1,000 normal displacements to 3,000 DOF.</u>
(1,000 displacements and 2,000 slopes)

2.1.1. Thermal Analysis.

Finite difference techniques are normally preferred by the heat transfer community for its analyses. This analysis is performed on a mathematical model with two hundred nodes, a moderately large model for finite difference heat transfer work. The heat transfer analysts are satisfied with the convergences and stability of their solutions and feel that they have accurately calculated the necessary temperatures.

2.1.2. Elastic Analysis.

This heat transfer model simulates an elastic system that will require fifteen hundred nodes in the structural model to adequately describe its behavior. To perform the thermo-elastic analysis in the structural model a temperature must be assigned to each of its fifteen hundred nodes. This requires an expansion of the data from the heat transfer analysis of two hundred nodes to the structural model of fifteen hundred nodes, an expansion of 7.5:1. This writer knows of no physical basis upon which this expansion may be performed. Typically a data base manager routine will be told to expand the data using a linear spatial interpolation routine between the two sets of nodal points to achieve a best first order fit of the data. The interpolation maps the two hundred temperatures onto a field of fifteen hundred points. The resulting temperature field does not necessarily represent a physically reasonable solution to the thermal problem at hand.

In elasticity an analyst normally faces several choices that affect the quality of his analysis: "modal" versus "direct" solution of the dynamic equations and "reduced" versus "unreduced" matrixes. "Unreduced" solution of the problem being discussed involves working with a matrix of order 9,000 (1,500 nodes with 6 degrees of freedom per node = 9,000 DOF) and finding solutions of it in either the static, time or frequency domains. In the static domain solutions of the order 9,000 may be routinely performed on contemporary mainframe computers and even work stations. However, in dynamics (either time or frequency domain) the solution of such a matrix is considered to be extravagant and the matrix is usually "reduced" before the dynamic analysis is performed. It is important to understand that although this reduction is sometimes desirable, it is by no means necessary for the solution since "direct" solution of the unreduced matrix is always available.

Several methods of matrix reduction are available but most of them are variants of Guyan reduction. Guyan reduction performs a rigorous reduction of both the stiffness and mass matrixes for a static analysis. However, for dynamic analysis it can be shown that the mass matrix is not properly reduced and therefore introduces errors in the dynamic calculation. The quality of the reduced matrixes is also affected by the selection of the points that are to remain in the reduced matrixes; an unwise selection of points will significantly skew the apparent behavior of the solution. Since there is no rule for reducing the mass matrix in dynamics problems it should be clear that even routines that "optimize" and "automate" the selection of the reduced matrix set are subject to errors in the reduced mass matrix of the structure.

Selecting eigenvalues and eigenvectors for use in subsequent solution routines is also a matrix reduction process. A system of 9000 equations representing 9000 degrees of freedom will have 9000 characteristic solutions (eigenvalue and eigenvector sets). By selecting twenty-five of these we are selecting the behavior associated with twenty-five degrees of freedom out of the complete set of 9000 (modal degrees of freedom). Modal solution routines are attractive because they use a small subset of the complete system. If the eigenvalues and eigenvectors are calculated from the complete set of equations they will represent accurate stiffness and mass matrixes for the selected degrees of freedom. If, however, the matrix was first reduced by a Guyan-like reduction technique (a common practice) the resulting eigenvalues and eigenvectors will suffer from the inaccuracies in reducing the mass matrix.

It should be appreciated then that by reducing the matrix size the analyst is also reducing the information available in the final analysis steps. The analyst must be constantly aware that he may not be including enough information in his reduced solution set of equations. Consider the case of designing a control system.

2.1.3. Control Systems Analysis.

Control system design software cannot calculate disturbances from temperature distribution data. Therefore, the effects of the heat loading on the structure must be estimated and converted to usable dynamic force, pressure or inertia loads. These estimated equivalent loads are used in conjunction with eigenvalues and eigenvectors from modal analysis to estimate the response of the instrument to the disturbance.

Structural analysis finite element routines generally calculate the lowest "n" eigenvalues (frequencies) where "n" is a positive integer. This is reasonable in stress analysis since it has been shown that the lowest frequencies tend to produce the highest stresses in the structure. However, consider a structure designed to contain active elements to control its geometry in dynamic environments. The structure, even though it may be relatively soft, will contain stiff mountings and load distribution members near the attachment of the actuators. The structure near the actuators will be very stiff, characteristic of high resonant frequencies. The set of "n" eigenvalues needs to capture the high frequency behavior of the structure near the actuators. If the set of eigenvalues is truncated at too low a frequency they will not be able to simulate the response of the structure to the actuator forces, causing sizeable errors between analysis and actual behavior. If "n" is large enough to include the structural stiffness in the vicinity of the actuators it may run from several hundred to several thousand eigenvalues; 1000 eigenvectors (in the above problem) represents about 300 megabytes of data.

It is clearly desirable to reduce the problem as much as possible but there is no automated way to be sure that the system has kept the information it needs to get accurate analytical answers. Control systems software is very good at solving the problems it is given and optimizing the control functions. But one must be aware that the control system design can be no better than the eigenvalues and eigenvectors that it is given.

2.1.4. Optical Analysis.

When the system performance is evaluated by data from either the finite element analysis or the control system analysis it is taken into an optical design code. Assuming that the finite element model uses the same number of points on each surface as the number of rays in the optical design it will only be necessary to format this data for the optical code by interpolating from the geometry of the structural model of each surface (usually rotational symmetry) to the geometry necessary for input to the optical code (usually rectangular coordinates).

The optical design code will further interpolate the input displacement data to determine the displacement at each ray intercept point in the system. It will also differentiate the interpolated input data to determine the slope changes (in two orthogonal directions) at each ray intercept; thus the optical design code generates 3,000 necessary pieces of data from the 1,000 normal displacements in the input file. These data are necessary for the optical code to perform its ray-trace function which depends upon the slope data as well as the normal displacement data.

Clearly the quality of the optical evaluation will depend upon both the quality of the input data generated by the data base manager from the finite element code output and the quality of the interpolation and differentiation algorithms used in the optics code. Evaluation of these effects has been discussed elsewhere (Hatheway, 1986) with the conclusion that,

> although seemingly reasonable estimates of the expected errors may be made they usually rest upon unsupported assumptions, the result being that the magnitudes of the interpolation and differentiation errors are unbounded. A worst case estimate of the errors in a specific problem may not be possible.

This method appears to work reasonably well on simple problems that are not too challenging. However, success on challenging problems with surface deformations represented by high order polynomials or Zernike terms have yet to be called to this writer's attention.

2.1.5. Summary of Integrated Analysis.

Integrated Analysis is the method of choice in the industry. It is a natural extension of the traditional (in the late twentieth century) methods in which each discipline uses the techniques in which it feels the most comfortable. Integrated Analysis has spawned a new group of "analysts" who support the central data base manager, write routines to interface codes, prepare the macro-files that actually control the Integrated Analysis process and generally keep the system running. Evaluation of the results of Integrated Analysis are judged subjectively based upon the complexity of the models, the advancement of the present design over designs previously analyzed, built and measured and the difficulty of the parameters being evaluated. Integrated Analysis and its intellectual forebears have produced some of the outstanding achievements of our century: the two hundred inch Hale telescope on Palomar Mountain, the Apollo moon landings, the Hubble Space Telescope and the ten meter Keck telescope in Hawaii.

2.2. Unified Analysis

Not withstanding the successes of Integrated Analysis, some workers in the field (Genberg, 1983, 1987; Hatheway, 1984; Paxson, 1988, 1989) have investigated the possibility of another approach, using a single finite element code to model all the

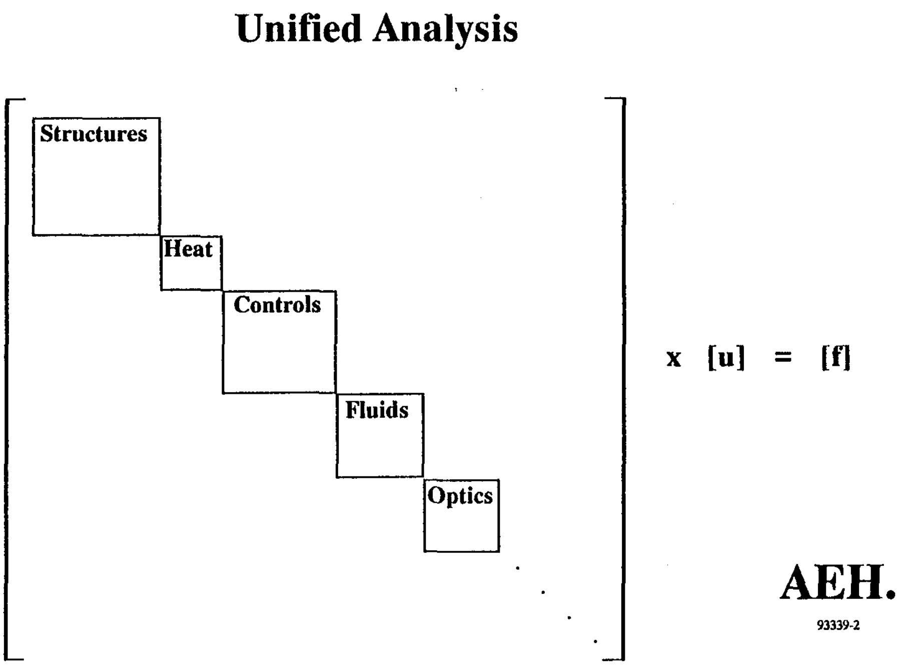

Figure 2. Unified Analysis prepares all of the disciplines for solution by the same software routine.

disciplines in a design. This technique invariably uses a structural analysis code (usually NASTRAN) as a computational vehicle. It is a challenge to the analyst to fit the equations of each discipline into the context of the finite element structural code (since the codes are dimensionless it is tempting to assume the are discipline-less as well). Since the lexicon of the code is structural mechanics, this always involves developing one or more "analogies" between the equations of structural mechanics and the equations of the discipline being modeled. When fully assembled the matrix equations of the system problem might be arranged as shown in Figure 2 which reads, in the language of structural mechanics, that the "stiffness matrix" multiplied by the "displacement vector" is equal to the "load vector." Let us call the procedure for solving these equations "Unified Analysis."

The unified method usually relies on linear, or piece-wise linear, equations throughout. This is a practical matter since many finite element codes can perform non-linear analyses when necessary but such analyses are very time consuming. The linearization of the equations provides very accurate solutions especially for optical systems, since the magnitude of the displacements is very small (Hatheway, 1988, SPIE).

The unified method has demonstrated its accuracy and utility in many applications. One example is a simple afocal telescope attachment (Figure 3) in which the method was able to predict both the bore sight error and the figure loss on the primary mirror and thereby help to correct a major problem in the telescope's construction (Hatheway, 1988, ASME). The unified model included analogies for reflection and interferometry in order to include them in the structural stiffness matrix. The model predicted the optical behavior under abnormal loading conditions.

Another example is a full scale (1:1) model of the Solar System (Figure 4). A constellation of high energy laser battle stations orbits the earth and intercepts "clouds" of re-entry vehicles near their mid-course apogee. The model includes active structures, adaptive optics, wave front sensors, acquisition and tracking sensors and a scoring system to assess the quality of the engagement, and the whole problem runs in less than an hour, nearly real-time. This unified model includes analogies for heat transfer, control systems, reflection, refraction, optical path difference and orbital mechanics in order to write these disciplines into the structural stiffness matrix. This model was developed in order to demonstrate the use of optical

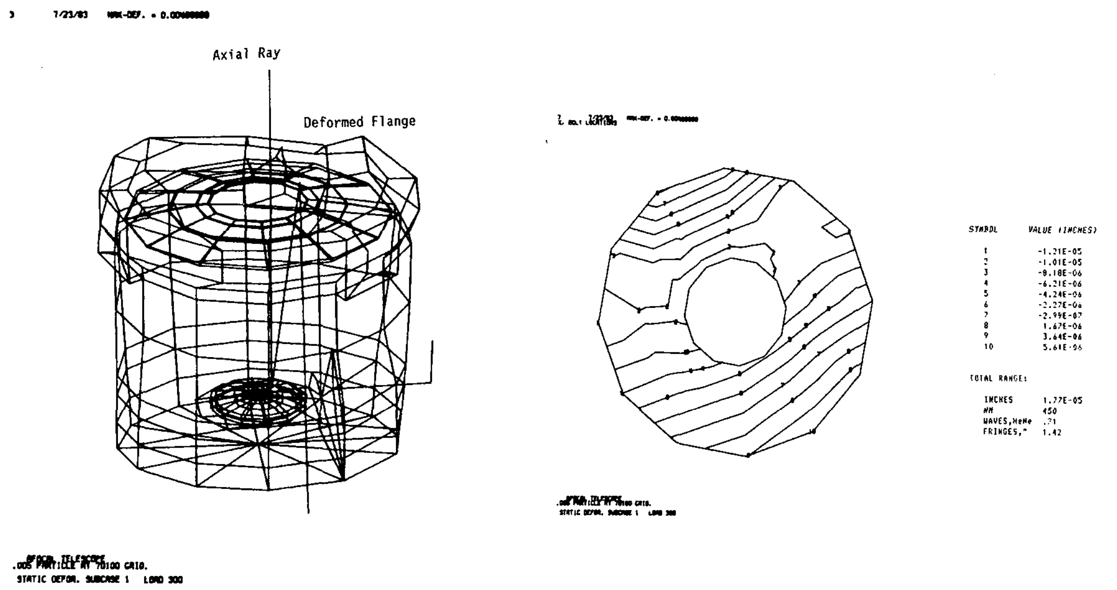

Figure 3. Reflection and interferometry in a Unified Analysis telescope model.

performance criteria in the optimization of active structures and adaptive optics, both of which are servo-elastic problems.

In Figure 4 the reader can see the Sun riding the Earth's limb while the Moon slides by just below. This analysis may seem like a video game simulation but it is an example of the realism (both subjective and quantitative) that can be achieved when applying analog techniques in the Unified Analysis method.

In other applications unified opto-mechanical models have demonstrated their ability to distinguish the various contributors to aberrations in optical systems; aberrations caused by non-uniform heating, gravitational sag, fluid pressures and transient accelerations have all been separated from the normal optical aberrations such as coma and astigmatism. The refraction analogy has been extended to include the dependence of the index of refraction on both temperature and strain states; systems with up to forty refractive surfaces have been successfully analyzed for the temperature and load effects on image quality. In all cases the Unified Analysis method has agreed well with both experimental results and conventional analytical results when available.

Models for Unified Analysis are constructed very meticulously in order to accommodate all the disciplines that need to be included. The structural model usually begins by modeling the rays or waves from an existing optical analysis in order to reflect the appropriate geometry of the intercept points and surfaces. The optical laws are then coded by analogy to the structural code. Finally the structural model (lenses, rings, mirrors, spacers, housings, etc.) is added using the optical geometry as a basis (Hatheway, 1988, ASME; 1988, SPIE). A little geometry and perhaps some elements may need to be added for heat transfer. Control system functions need to be added as well but usually no new geometry is required (the functions being entirely scalar and the forces acting on existing structural geometry).

During each phase of development of a Unified Analysis model the analyst needs to perform sufficient checks on the model to be sure that each discipline is accurately represented in the final product. Multi-discipline models which take advantage of analogies are particularly susceptible to error in dimensionality: the units that are convenient for structural mechanics are not necessarily convenient for optics or heat transfer so that some compromise must be made in the units that are used and some of the disciplines (perhaps all) may need to work in awkward or unconventional systems of units. The units employed

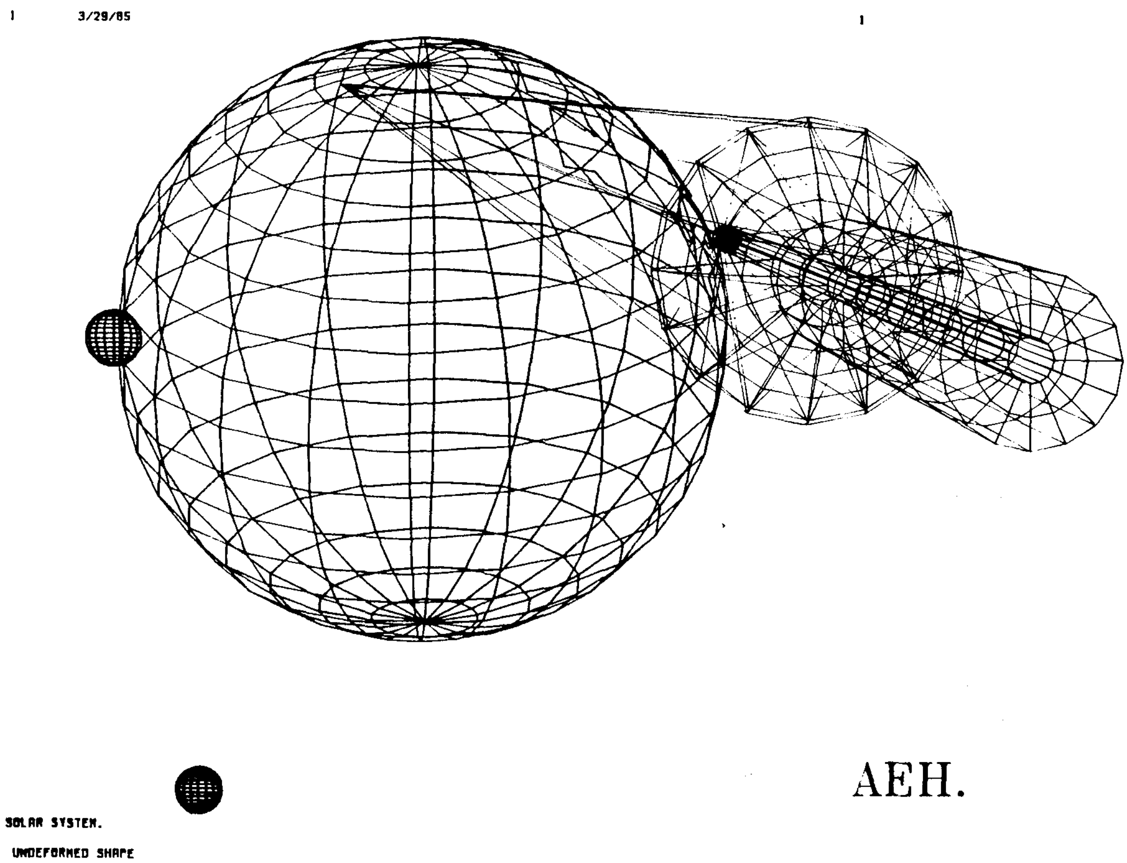

Figure 4. A Unified Analysis model of a spacecraft orbiting the Earth. The Sun and Moon are also shown (full scale).

in each discipline need to be checked and re-checked but finally running a simple problem with a known (or estimated) answer is the best test for the quality of the unified model. This process of "validation" not only debugs, corrects and extends the model, it also builds confidence in all participants in the activity, as they see the model run and produce meaningful results. The model may also be run to solve the complete set of equations without reduction or simplification by using direct solution methods in both the time and frequency domains.

Although Unified Analysis has demonstrated its performance (Hatheway, 1987) it still has two major handicaps: first, all disciplines need to agree to use the same code. Anyone who has ever attempted to get two analysts to agree on this issue will understand the difficulty of getting four or five departments (disciplines) in a large company to agree. The number of codes capable of Unified Analysis is very limited and they tend to be among the least user friendly so that few of the participants may understand the codes and the analogies that are used to unify the disciplines.

Second, linear equations are generally made to apply to the problem solution. Some finite element structural codes can accommodate non-linear analyses but they are not necessarily adapted to the types of non-linearities that are common in other disciplines such as optics and control systems. In optical systems this linearization is performed for the nominal optical design based upon data supplied by the optical designer. Since the purpose of the Unified Analysis is to investigate the magnitude of perturbations about the nominal design and the perturbations (displacements, temperature changes, index changes, etc.) are very small the linearized equations are capable of great accuracy (Hatheway, 1988). The accuracy of the linearization process may be calculated for each instrument. It depends upon the size of the instrument, the wave length of interest and to some extent on the angles of incidence and the refractive indexes.

Many optical instruments that use control systems to maintain the alignment and the surface figure rely on non-linear transfer functions that must cover a very large range of magnitudes in the perturbations they accommodate (sometimes approaching the dimension of the instrument itself). For this reason there is reluctance to model control systems in the linear finite element codes. However, during operation the instrument must remain in nearly perfect operating alignment and figure so that the operating perturbations must be very small (equivalent to less than about a quarter wave for diffraction limited systems). Therefore the same rationale used for linearizing the optical equations applies to the control system equations: the perturbations are very small so the linearized behavior will be a very accurate representation of reality. If any doubt exists about the accuracy available from an Unified Analysis, all the errors in each discipline may be explicitly calculated before preparing the model or running the computer. The analogies are quantitative and explicit.

3. THE CHALLENGES OF THE 21ST CENTURY.

A number of interesting astronomical observatories are being propose for the early part of the coming century (ASTRO-TECH 21, 1991):

1) The "Next Generation Space Telescope" is proposed to enter service about year 2010. It will be constructed on the surface of the Moon with an anticipated mission duration of fifteen years. The primary mirror will be between ten and sixteen meters diameter (hexagonal segments) and will operate at wave lengths between 100 nanometers and ten microns.

2) The "LAGOS" instrument will detect gravity waves in space using an interferometer operating at one and six-tenths microns. Its various portions are stabilized in solar orbit and the baseline length is 10^7 kilometers. Its performance (and success) will be determined by the thermo-elastic stability of the receivers and transmitters (equivalent to thermal stability of thousandths of a degree Kelvin over the ten year life). Its in-service date is about the year 2015.

3) Another very intriguing instrument is the "Filled Fizeau Telescope." It has a thirty meter "cross" dilute aperture and operates between 100 nanometers and one micron. The initial service date was not indicated but it can be assumed to be in the vicinity of the year 2020.

These and other programs will rely heavily on active control of structural alignment and figure for proper operation. They are large dimension (aperture or baseline) instruments and may operate in the ultraviolet and visible portions of the spectrum as well as the infra-red and sub-millimeter portions. Configuring these systems will be a very challenging task requiring the very finest analytical techniques and a very sharp understanding of the disciplines that make the systems work. If one thing is clear about the future it is that the same pressures that forced the expert panel method (experts at a conference table) to become the Integrated Analysis method (computer and data base manager) will persist for the indefinite future. Those of us who plan to build these instruments have our work cut out for us.

Integrated analysis will still be used throughout the industry because of its ease, convenience and historical successes. It is not clear however that it can be relied upon to analyze the more difficult problems, especially ones involving active or passive surface deformations. It is fundamentally limited by the physical reasoning (algorithms) in the transformations used in moving data from one code to another. Optical codes are further limited by only considering surface displacement components that are locally normal to the surface contour, thereby requiring the differentiation of a function that probably does not represent the behavior of the feature under consideration.

Unified methods will also continue to be used where circumstances permit but will probably not be accepted as the preferred tool for analysis of these systems (although it has a firm physical and analytical foundation its techniques are too far removed from the traditional methods of each of the disciplines). The current codes are not user friendly and are very time consuming to learn. The documentation is completely couched in the lexicon of structural mechanics which makes it even more difficult for outsiders to find out what they need to know.

It seems to this writer that a new analytical basis might be developed that allows us to model all the disciplines rigorously (without analogies) using the same data. This new basis should comfortably accommodate all the disciplines. In fact all disciplines must contribute to its development for it to be widely received and utilized. It will be comprised of both software routines and high capacity computers. New numerical methods (especially in electromagnetics) may reduce the demands on both the software and the computers, the P-method of finite element analysis is a promising example of what might be done but much more work is needed.

This new basis for analysis would have for its purpose the validation of all the design and development analyses that have been performed during the design of a particular instrument. Its architecture should be tailored to the solution of extremely large problems in both statics and dynamics without reduction or simplification. It should provide accurate replication of test data when such data become available and would become the final audit of our grasp of the mechanics of the particular instrument. It may only be run a few times in the course of a project but it would provide a basic "calibration" capability for each of the disciplines and the individual computational methods they choose to use during design.

Considering the ingenuity with which astronomers approach their science, it seems to this writer that the structural mechanics community might contemplate improvements in it its ability to serve them. With this common goal in mind he offers the following possibilities:

Structural mechanics finite element analysis codes ought to become fully kinematic codes. Presently they derive just enough kinematic (displacement) information to calculate one characteristic state of stress for each element. Some rotational information is not derived because it is not needed to calculate the stress. This rotational information is needed by other disciplines (especially optics) even though they have learned to get along without it so far. Integrated Analysis suffers particularly from this deficiency. Analysts who perform Unified Analysis are usually able to "fool" the code into providing these displacements on a semi-rational basis because they need them to complete their job.

The finite element method should finally out-grow the structural mechanics discipline. The method is applicable to many other fields of mechanics and we are just beginning to see some commercial development of codes for heat transfer, and electro-magnetics. This recent activity should be expanded so all disciplines can accept its capabilities (and be familiar with its limitations). If analysts in the disciplines are unwilling to use the same code, at least they need to know each others' codes well enough to be able to knowledgeably provide them input data.

Integrated Analysis needs a fundamental physical foundation for the data expansions, reductions, and interpolations that are performed. Without this physical foundation we will not be able to trust the analytical results and the disciplines will have to be very conservative to avoid disgrace.

Computational capacity and speed must grow dramatically. One way to linearize the optical equations is to use the wave equations themselves rather that ray tracing. A one meter class instrument operating in the visible (say 500 nanometers) and modeled in an electromagnetic wave finite element code using todays modeling strategies would require 10^{21} nodes, a monumentally large number by today's practices (10^4 or 10^5). However, a machine capable of processing such a matrix will be able to truly unify the critical fields of mechanics without artifices and simplifications. Numerical method and software developments may reduce the matrix size by several orders of magnitude (as in P-method analysis) but the number of equations to be solved will still be larger than anything currently contemplated. Computer scientists have talked about developing high capacity machines (optical computing, parallel processing, etc.). Well, here is a practical requirement that is not made entirely out of straw. Given an example of an immanent need and some cultivation and encouragement, who knows what kind of computational engines may be available in twenty years to inspire the subsequent generations of designers?

4. REFERENCES

ASTROTECH 21, Optics Technology Workshop, Pasadena, CA, March 1991.

Bathe, Klaus-Jürgen; Wilson, Edward L. Numerical Methods in Finite Element Analysis, Englewood Cliffs, New Jersey, Prentice- Hall, 1976, pp. xiii-xv.

Genberg, Victor; Oinen, Don; Fronheiser, Sharon. "Diffuse Illumination with MSC/NASTRAN," MSC/

NASTRAN Users' Conference Proceedings, The MacNeal-Schwendler Corporation, Los Angeles, CA, March, 1983.

Genberg, Victor. "Optical pathlength calculations via finite elements," Structural Mechanics of Optical Systems II, Volume 748, SPIE, January, 1987.

Hatheway, Alson E. "Finite element methods for evaluating optical system performance," Optical System Engineering IV, Volume 518, SPIE, October, 1984.

Hatheway, Alson E. "A Modeling Strategy for Analysis of Optomechanical Systems," MSC/NASTRAN Users' Conference Proceedings, The MacNeal-Schwendler Corporation, Los Angeles, CA, March, 1986.

Hatheway, Alson E. "Optomechanical analysis strategies," Structural Mechanics of Optical Systems II, Volume 748, SPIE, January, 1987.

Hatheway, Alson E. "Optical simulations in the finite element method: an overview," Simulation and Modeling of Optical Systems, Volume 892, SPIE, January, 1988.

Hatheway, Alson E. "Optics in the Finite Element Domain," Computers in Engineering 1988 - Volume 3, Book No. G0431C, American Society of Mechanical Engineers (ASME), New York, 1988

Love, A. E. H. A Treatise on the Mathematical Theory of Elasticity, 4th. ed., 1927, New York, Dover Books (1944), pp. 1-31.

Paxson, Ernest B., Jr. "Simulation of Room Temperature Control System (RTCS) with MSC/NASTRAN," The 1988 MSC World Users' Conference Proceedings, The MacNeal-Schwendler Corporation, Los Angeles, CA, March, 1988.

Paxson, Ernest B., Jr. "Simulations of Small Structures-Optics- Controls Systems with MSC/NASTRAN," The 1989 MSC World Users' Conference Proceedings, The MacNeal-Schwendler Corporation, Los Angeles, CA, March, 1989.