Science & Technology Facilities Council
Rutherford Appleton Laboratory

Development and history of sparse direct methods

Iain S. Duff

iain.duff@stfc.ac.uk

STFC Rutherford Appleton Laboratory

Oxfordshire, UK.

and

CERFACS, Toulouse, France

Homepage: http://www.numerical.rl.ac.uk/people/isd/isd.html

Outline

- Ancient times

Outline

- Ancient times
- The birth

Outline

- Ancient times
- The birth
- Infancy

Outline

- Ancient times
- The birth
- Infancy
- Turbulent teens

Outline

- Ancient times
- The birth
- Infancy
- Turbulent teens
- Maturity

Outline

- Ancient times
- The birth
- Infancy
- Turbulent teens
- Maturity
- Senility?

Direct methods

Gaussian Elimination

$$\mathbf{PAQ} \rightarrow \mathbf{LU}$$

Permutations $\mathbf{P}$ and $\mathbf{Q}$ chosen to preserve sparsity and maintain stability

$\mathbf{L}$: Lower triangular (sparse)
$\mathbf{U}$: Upper triangular (sparse)

SOLVE:

$$\mathbf{Ax} = \mathbf{b}$$

by

$$\mathbf{Ly} = \mathbf{Pb}$$

then

$$\mathbf{UQ^T x} = \mathbf{y}$$

Ancient times

Large systems were solved in ancient times by iterative methods, principally relaxation methods (Richardson, Southwell, Fox)

and using SOR and related techniques ... thesis of David Young (1950), book by Varga (1962).

The earliest references from standard sparse matrix books date to the early 1950s but these papers are principally concerned with graph theory and combinatorics and don’t really reference or use sparse matrices or even sparse data structures (König, Woodbury, Hall).

Ancient times

16. Experiments on the Inversion of a 16X16 Matrix[1]

John Todd*

Introduction

Some experiments have been carried out in the Computation Laboratory of the National Bureau of Standards on the inversion of a certain 16×16 matrix, using the following three methods: (1) G. W. Petrie's arrangement of the Gauss elimination process,[2] (2) a Monte Carlo process,[3] and (3) an iteration method.[3] It is the purpose of this note to describe and compare the results obtained.

The matrix was a 16×16 matrix of the so-called Leontief type [1, 2],[4] representing certain inter-industry relations. A matrix of the same type, but of order 40×40, has been investigated by J. L. Holley of the Air Comptroller's Office, USAF; in particular, it has been inverted, using the UNIVAC.

The actual matrix inverted was $B=I-A$, where $10^5 A$ is:

0	36265	0	0	0	0	0	0	4646	3169	8939	0	421	0	1355	0
5158	0	0	0	0	0	0	0	1382	103	2466	0	120	0	6995	5
188	0	0	9779	12795	2679	0	0	0	1062	0	0	0	909	0	5
20	38	77	0	30	128	145	264	29	85	99	335	0	0	5	0
69	2225	424	10425	0	128	0	275	882	874	83	0	601	2434	253	104
49	173	4165	2398	3252	0	145	550	2705	86	17	0	3065	0	0	30
128	913	1080	1209	696	319	0	66	3528	754	33	561	300	174	200	10
3794	1109	12225	1835	2087	3890	8475	0	5322	2861	1986	2578	2464	8653	2822	277
2915	905	1388	709	1361	638	1065	286	0	1867	5314	3475	2404	201	0	173
869	1682	77	709	907	383	1937	22	1764	0	761	448	1563	134	1194	6747
524	324	0	1501	121	64	97	0	382	1388	0	7960	1202	0	146	119
326	23	0	3878	242	0	0	11	29	34	463	0	60	348	58	139
10	0	0	250	136	0	0	0	0	137	1639	0	0	321	331	1288
7845	2942	10297	2252	1996	4719	14286	24210	6616	6630	844	3475	421	0	19	0
3537	2527	926	1877	1165	510	1840	385	3323	2895	3145	897	2764	642	0	515
5444	2821	501	1772	1119	191	630	429	5410	1079	1374	4596	962	976	5294	0

Method 1

The actual running time was some 8 hours on the 604 Calculating Punch. The error in the check sums carried was about 160 units in the last (eighth) place. The resultant matrix is denoted by G and is available.

*National Bureau of Standards.
[1] This work has been supported by the Air Comptroller's Office, USAF.
[2] See the previous paper. The IBM operations were carried out under the direction of Helen V. Hammar of the NBS Computation Laboratory.
[3] These were carried out on SEAC under the direction of Karl Goldberg of the NBS Computation Laboratory.
[4] See also [3] for description of the inversion of a 38X38 matrix of this type, by a Gaussian process, on the Aiken Relay Computer, Mk. II.

Ancient times

One of the first papers to directly link graphs with sparse elimination was due to Seymour Parter in 1961.

Major application areas were the simplex method and backward difference formulae for solving stiff ODEs.

Other areas strongly represented in ancient times were:

Power systems (Bill Tinney)

Electrical networks and circuit design (Bob Brayton, Gary Hachtel, Gabriel Kron, Alberto Sangiovanni-Vincintelli, Donald Steward)

Chemical engineering (Roger Sargent and Art Westerberg)

Stiff ODEs

Explicit methods for solving ordinary differential equations had particular difficulty when the problem was stiff.

The use of implicit methods based on backward difference formula was brought to the notice of linear algebraists by the landmark talk by Bill Gear at the IFIP Conference in Edinburgh in 1968.

The equations that were solved in Gear's backward difference formula were of the form

$$(I - \alpha h J)x = b$$

where J was the Jacobian, h the stepsize, and α a parameter.

These equations could be very evil but very sparse.

Liniger at IBM Yorktown and Curtis at Harwell both had equations of this kind to solve.

Linear Programming

The linear programming community was one of the first to embrace sparsity

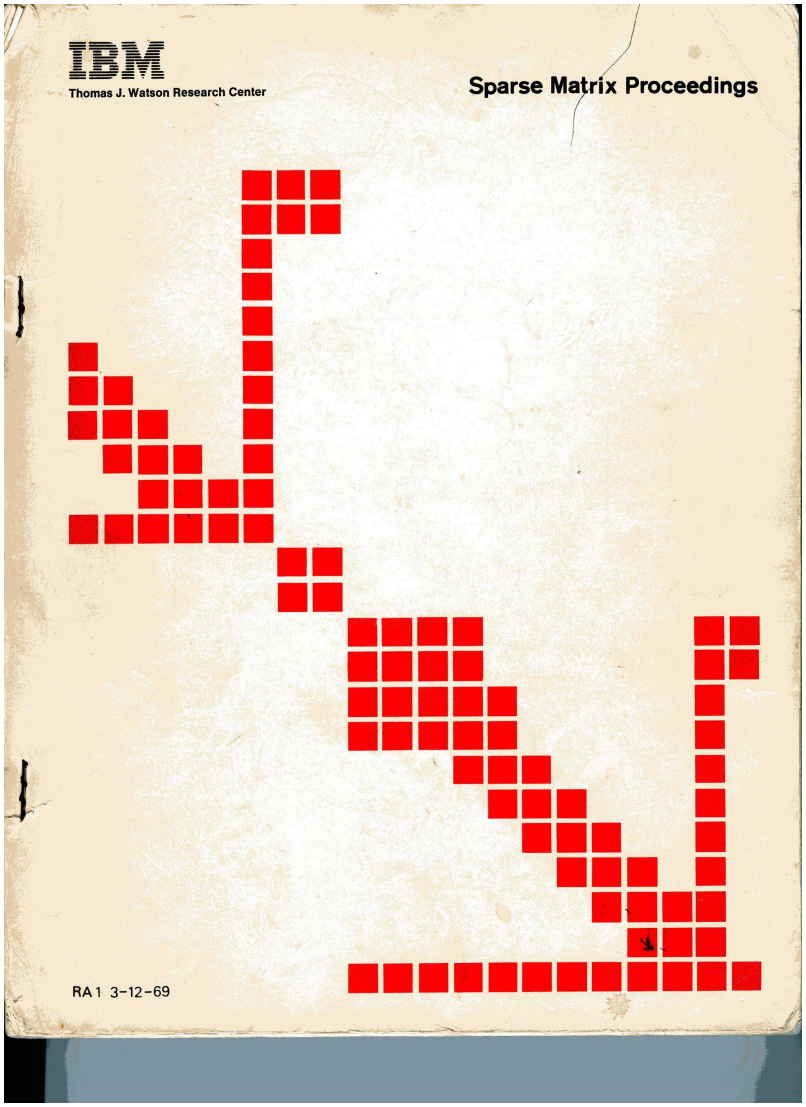
- 1947 The simplex method. George Dantzig
- 1954 The product form of the inverse. Dantzig and Orchard-Hays
- 1955-56 Orchard-Hays implements sparse LP codes on the Johnniac at RAND
- 1957 The elimination form of the inverse. Harry Markowitz
- 1961 Martin Beale doing sparse programming at CEIR (UK) later Scientific Control (SCICON)

Ancient times

Harry M. Markowitz

The Birth

The Birth



Iain Duff.
Oxford Computing Laboratory.

Proceedings of the Symposium on Sparse Matrices and Their Applications

Held at the IBM Watson Research Center September 9-10, 1968

Organized and Sponsored by the Mathematical Sciences Department

Ralph A. Willoughby, Editor

International Business Machines Corporation
Thomas J. Watson Research Center
Post Office Box 218
Yorktown Heights, New York 10598

RA 1 (# 11707)
March 12, 1969
Numerical analysis

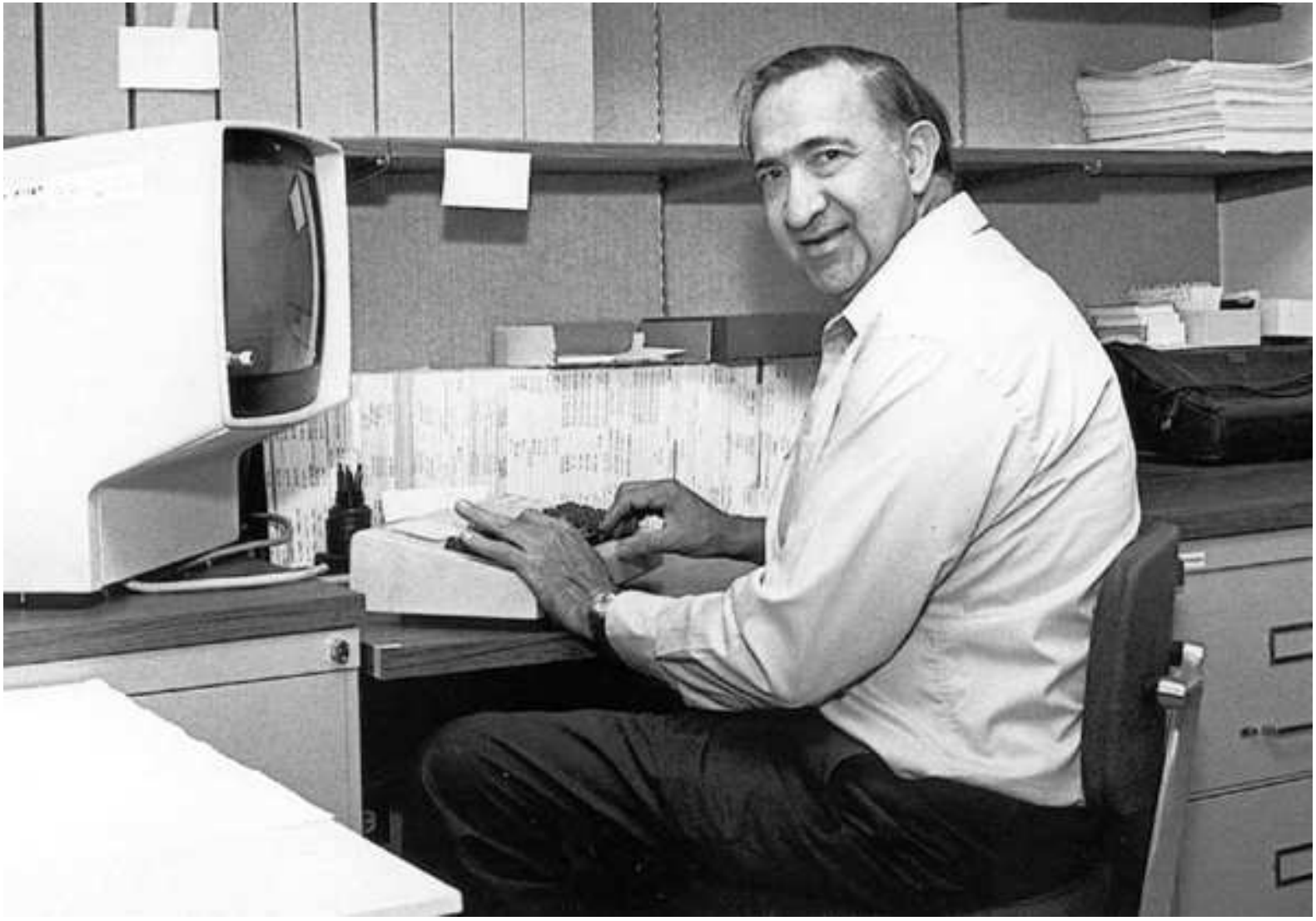
The Birth

The first Sparse Matrix Symposium was organized at IBM Yorktown Heights in 1968 by staff of the Mathematical Sciences Department: Robert A. Brayton, Fred G. Gustavson, Alan J. Hoffman, and Philip S. Wolfe, and Ralph A. Willoughby, who edited the Proceedings.

I have always thought of Ralph as the “father” of sparse matrix technology.

Obituary written by Jane Cullum in SIAM News 35 (7), September 2003.

The Birth



Ralph Willoughby 1923 - 2001 [courtesy of family and SIAM]

The Birth

Introduction

by Ralph A. Willoughby

A symposium on sparse matrices and their applications was held at the Thomas J. Watson Research Center on September 9-10, 1968. The meeting was organized by Robert K. Brayton, Fred G. Gustavson, Alan J. Hoffman, and Philip S. Wolfe, who are members of the Mathematical Sciences Department at IBM Research. A total of 124 people registered for the meeting: 81 were non-IBMers from industry, government agencies, and universities; 25 were from other IBM locations; and 18 were from the IBM Watson Research Center.

The first day was devoted to a discussion of basic techniques and recent advances. On the second day, applications of sparse matrix techniques were emphasized. A panel discussion on new or needed work and open questions was held at the end of the second day. In this Proceedings we are publishing extended abstracts of the talks and also an edited version of the panel discussion.

The desirability of such a meeting was realized as a result of a number of research investigations carried on by members of the Mathematical Sciences Department. This research concerned the use of direct methods for arbitrarily sparse matrices in solving systems of linear equations in problems of applied mathematics.

In searching the literature for references on the use of sparse direct methods, a number of things became clear. For one thing, in the general numerical analysis periodicals and books, direct methods have usually been considered only for small full matrices, while large sparse matrices were handled by iterative techniques. On the other hand, in a number of quite different application areas, sparse direct methods have been extensively developed and made a part of programming packages.

The Birth

CONTENTS

*presented talk

The Birth

Session II -- W. Givens, Chairman

Special Evening Session -- R. K. Brayton, Chairman

The Birth

APPLICATIONS

Session III -- J. R. Edmonds, Chairman

The Birth

Session IV -- A. J. Hoffman, Chairman

The Birth

A number of “currently hot" topics were discussed at the 1968 meeting!

- Hybrid methods
- MC64 scaling before Olschowka and Neumaier (1996)
- Use of single precision arithmetic to get full accuracy
- Permutations to block triangular and bordered block diagonal form and “tearing” in general
- Modified Markowitz

The Birth

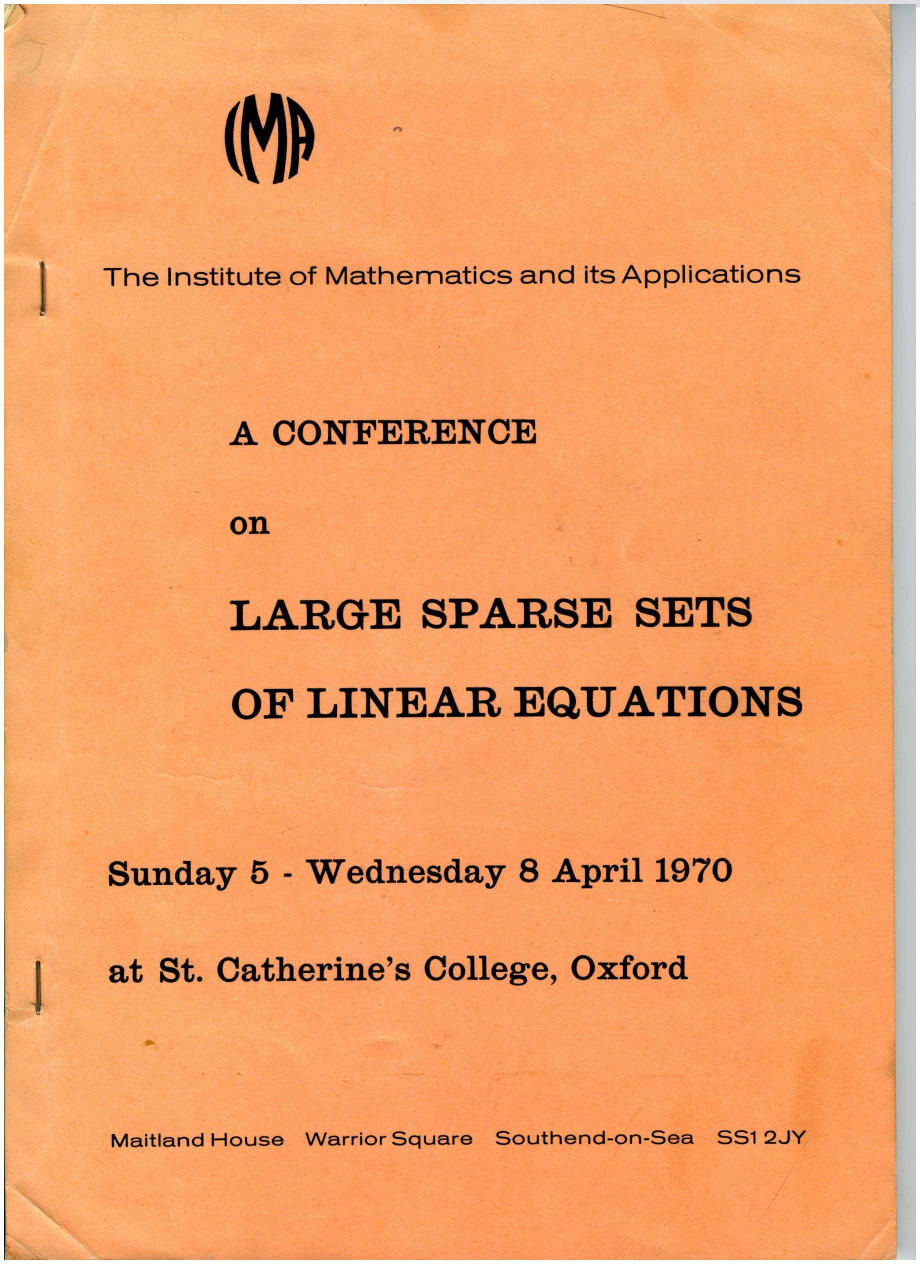
At this time perhaps the only “general purpose” code was that of Fred Gustavson at IBM Yorktown who used a generated code approach for solving problems coming from stiff ODEs. This code was later called GNSOIN.

Papers by Brayton, Gustavson, and Willoughby (1970), Gustavson, Liniger, and Willoughby on GNSOIN (1970), and by Hachtel, Brayton, and Gustavson on variability typing (1971).

Shortly afterwards, Curtis and Reid developed codes MA17 (symmetric positive definite) and MA18 (unsymmetric), the latter for the same application as above.

The Christening

The Christening



(IMA

The Institute of Mathematics and its Applications

A CONFERENCE

on

LARGE SPARSE SETS

OF LINEAR EQUATIONS

Sunday 5 - Wednesday 8 April 1970

at St. Catherine's College, Oxford

Maitland House Warrior Square Southend-on-Sea SS1 2JY

The infancy

Around this time, the first theses on sparse direct methods appeared.

Donald Rose	1970	Harvard	Symmetric elimination on sparse positive definite systems and the potential flow network problem
Alan George	1971	Stanford	Computer implementation of the finite-element method
Iain Duff	1972	Oxford	Analysis of sparse systems
Andrew Sherman	1975	Yale	On the efficient solution of sparse systems of linear and non-linear equations

The infancy

By the mid 1970s, there were only really three widely available sparse matrix packages: HSL, SPARSPAK, and YSMP.

Harwell Subroutine Library (HSL)

Main reason for developing these codes was to enable Alan Curtis to solve stiff ODE problems. First codes were MA17 (symmetric positive definite) and MA18 (unsymmetric) by Curtis and Reid in 1971 and then MA28 (unsymmetric) by Duff in 1977 and MA27 (symmetric) by Duff and Reid in 1982.

YSMP (1975) LLL User's Guide by Andrew Sherman. Paper by Eisenstat, Gursky, Schultz, and Sherman in 1982

SPARSPAK (1978) University of Waterloo. George and Liu (then Ng).

By 1982 a sparse software catalogue also lists codes from IBM, Bell Labs, CRAY, FPS, SSLEST and Y12M from DTU, COSMIC, NSPFAC, ...

Maturity

SPARSE BOOKS

1973	Tewarson	Sparse Matrices
1976	Brameller, Allan and Hamam	Sparsity
1981	George and Liu	Computer Solution of Large Sparse Positive Definite Systems
1983	Østerby and Zlatev	Direct Methods for Sparse Matrices
1984	Pissanetsky	Sparse Matrix Technology
1986	Duff, Erisman and Reid	Direct Methods for Sparse Matrices
1991	Zlatev	Computational Methods for General Sparse Matrices
2006	Davis	Direct Methods for Sparse Linear Systems

Maturity

Conference Proceedings

1969	Willoughby	Sparse Matrix Proceedings
1971	Reid	Large Sparse Sets of Linear Equations
1972	Rose and Willoughby	Sparse Matrices and their Applications
1973	Himmelblau	Decomposition of Large-Scale Problems
1976	Bunch and Rose	Sparse Matrix Computations
1977	Barker	Sparse Matrix Techniques
1979	Duff and Stewart	Sparse Matrix Proceedings 1978
1981	Duff	Sparse Matrices and their Uses
1985	Evans	Sparsity and its Applications

Maturity

Other publications and meetings that established the field.

- 1975 Symposium on Sparse Matrix Computations. Argonne
- 1977 A Survey of Sparse Matrix Research. Proc IEEE. Duff
- 1980 Direct Methods for solving large sparse systems. SIAM News. George
- 1982 Sparse Matrix Symposium 1982. Fairfield Glade, Tennessee.
- 1989 SIAM Symposium on Sparse Matrices. Salishan, Oregon
- 1990 Sparsity in Large Scientific Computations. IBM Europe Institute. Oberlech, Austria.
- 1995 International Linear Algebra Year at CERFACS. Direct methods workshop 1995. BIT **37**, 1997.
- 1996 SIAM Meeting on Sparse Matrices. Coeur d'Alene, Idaho.
- A number of CSC (Combinatorial Scientific Computing) Meetings. Roughly biennial.

Turbulent times

Gordon Research Conference on Flow in Permeable Media. Proctor Academy, Andover, NH. July 30 - Aug 3, 1984

Turbulent times

Gordon Research Conference on Flow in Permeable Media. Proctor Academy, Andover, NH. July 30 - Aug 3, 1984

Talk of ISD

The solution of large sparse asymmetric matrices by direct methods.

Turbulent times

Gordon Research Conference on Flow in Permeable Media. Proctor Academy, Andover, NH. July 30 - Aug 3, 1984

Talk of ISD

The solution of large sparse asymmetric matrices by direct methods.

Punchline

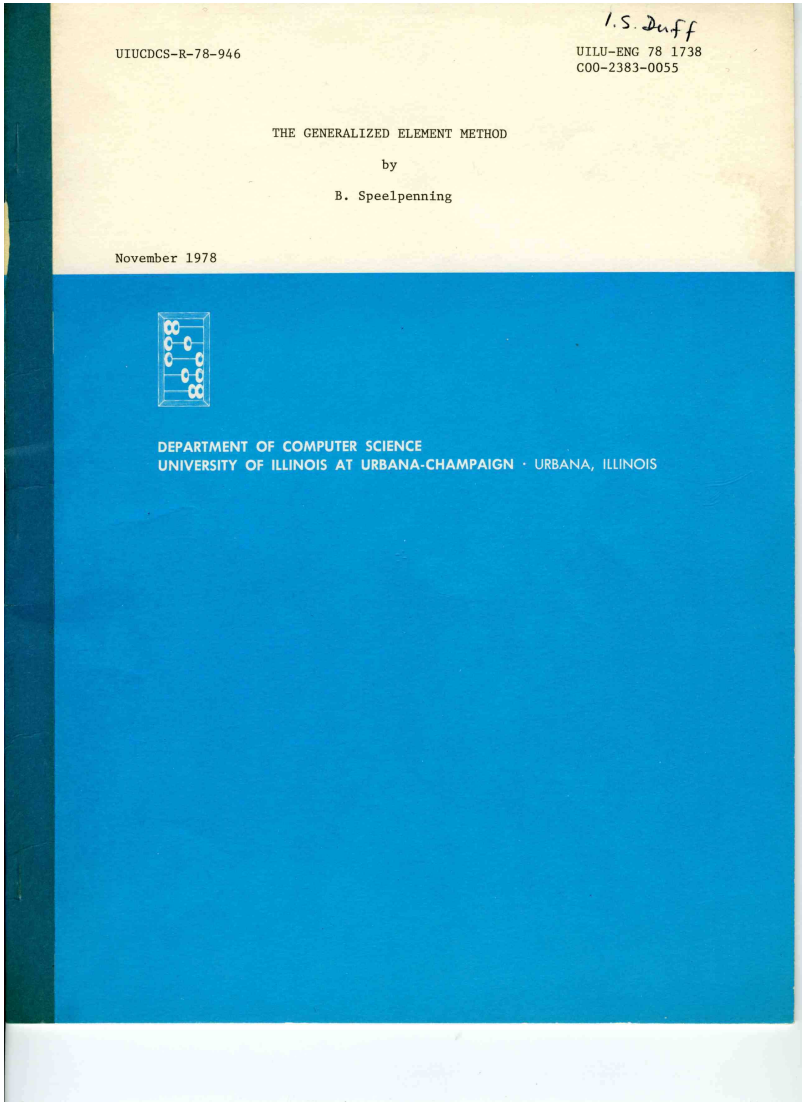
Real men use direct methods

Maturity

Grid dimensions	Matrix order	Work to factorize	Factor storage
$k \times k$	k^2	k^3	$k^2 \log k$
$k \times k \times k$	k^3	k^6	k^4

$\mathcal{O}$ complexity of direct method on 2D and 3D grids.

Maturity ... multifrontal



I.S. Duff

UIUCDCS-R-78-946

UILU-ENG 78 1738
C00-2383-0055

THE GENERALIZED ELEMENT METHOD

by

B. Speelpenning

November 1978

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · URBANA, ILLINOIS

Speelpenning's contribution .. 1973

- Background is very much in frontal methods (Irons 1970) and finite-element computations
- Makes break from wavefront/frontal ordering
- Uses generalized element formulation
- Combines elements on elimination to form larger elements
- Does multiple eliminations within a front
- No numerical experiments
- Preamble to 1978 reprint of 1973 report references use in DIANA system at TNO-IBBC for solving systems from structural analysis of order 5000 with "bandwidth" about 400

Early multifrontal developments

- Sherman (1975) : compressed storage scheme
- Eisenstat, Schultz, and Sherman (1976) : element merge tree
- Peters (1980) : substructuring (claims not generalized element)
- George and Liu (1980) : generalized element approach for minimum degree
- Duff and Reid (1983) : paper on background to multifrontal code MA27
- Reid (1984) : TREESOLV (out-of-core)
- Duff (1986) : prototype parallel implementation on Alliant FX/8
- Liu (1992) : detailed description of multifrontal method

MA27 ... 1982

Main contributions

- Full implementation of tree for all phases (including node amalgamation)
- Use of stack for intermediate frontal matrices
- Symmetric indefinite factorization using Duff, Reid, Munksgaard, and Nielsen's (1979) sparse adaptation of a variant of two-by-two pivoting strategy of Bunch and Parlett (1971). This was perhaps the first implementation of rook pivoting that Saunders used later in LUSOL.
- Exploitation of vector machines (pre-dated higher-level BLAS) using experience from frontal solvers (MA32)

Note that the only codes with which we could then compare MA27 were MA17, YSMP and SPARSPAK (plus a code of Munksgaard)

Senility?

Senility?

HYBRID METHODS

COMBINING DIRECT AND ITERATIVE METHODS

(can be thought of as sophisticated preconditioning)

Multigrid

Using direct method as coarse grid solver.

Domain Decomposition

Using direct method on local subdomains and “direct” preconditioner on interface.

Block Iterative Methods

Direct solver on sub-blocks.

Partial factorization as preconditioner

Factorization of nearby problem as a preconditioner

Conclusions

Sparse Direct Methods have had a short, distinguished, and very active history

There are still many future research challenges

A range of techniques involving both sparse direct and a range of sparse iterative solvers is required to solve the really challenging problems of the future

Conclusions

THANK YOU

for your attention