Moore's Law

- Every 18 months, the speed of your computer is doubled

- Every 18 months, the memory on your computer is doubled

- At the same time, the cost of your computer goes down - not quite exponentially, because the box does not become much cheaper!

- A good number to look at

$$R_{1970} = \frac{Cost\ of\ CPU\ time}{Cost\ of\ human\ time}$$

- 1970 is the year

- Different CPUs, different humans, etc.

Observation

- $R_{1945} >> 1000$
- $R_{1960} >> 100$
- $R_{1970} >> 10$
- $R_{1980} \sim 1$
- $R_{2000} << 0.01$
- Unlike men, not all CPUs are created equal! But then, most CPUs do not vote...
- The thing is not slowing down, though eventually . . .
- What should we be doing as applied mathematicians, numerical analysts, etc.?

Consequences

- Ticket reservations
- Phone systems
- Tactical bombing
- Experimental science
- Manufacturing
- . . .

Missing from the list

- Philosophy
- Theater
- Politics
- Dealing with teen-age children
- Mathematics
- Numerical simulation of physical phenomena (???!!!)

Structure of the Talk

- Changing paradigm in the numerical use of computers
- Interaction of Moore's law with numerical algorithms
- Characteristics of a modern numerical algorithm
- Example: rapid evaluation of radiation fields
- Pontification

Paradigm as of 1945

- Critical mission (Manhattan project, for example)
- Willingness to expend human time on programming (ouch!), debugging of the numerical scheme, interpretation
- Limited computer resources: only small-scale problems can be solved
- Extremely uncomfortable programming environment
- Air of heroism and desperation
- No difference between theoretical numerical analysts and practitioners
- Numerical approaches appropriate to small-scale problems
- Numerical algorithms usually written from scratch

Paradigm as of 1970

- Mission not necessarily critical (oil exploration, NACA airfoils, more involved airdynamics, civil and mechanical engineering, rocket fuel stoichiometry, . . .)
- Willingness to expend human time on programming (still pretty uncomfortable), interpretation
- Improved computer capabilities; CPU time still quite expensive, but the flop rate is much higher; one can try running things at night
- The air much less heroic; most applications in non-desperate environments
- Numerical algorithms appropriate to small-scale problems
- Most numerical codes written from scratch

Paradigm as of 2000

- Mission usually not critical: computer games, medical imaging, design of fishing rods, Boeing-767's . . .

- Limited willingness to expend human time on programming (could be fun, though!), interpretation. . . and most interpreters are not named Teller, Ulam, or Fermi. . .

- Very much improved computer capabilities; CPU time dirt cheap, and flop rate is about to become gigaflop rate

- Air not heroic at all; lots of applications, and most in non-desperate environments

- Numerical algorithms appropriate to small-scale problems

- Most numerical codes written from scratch

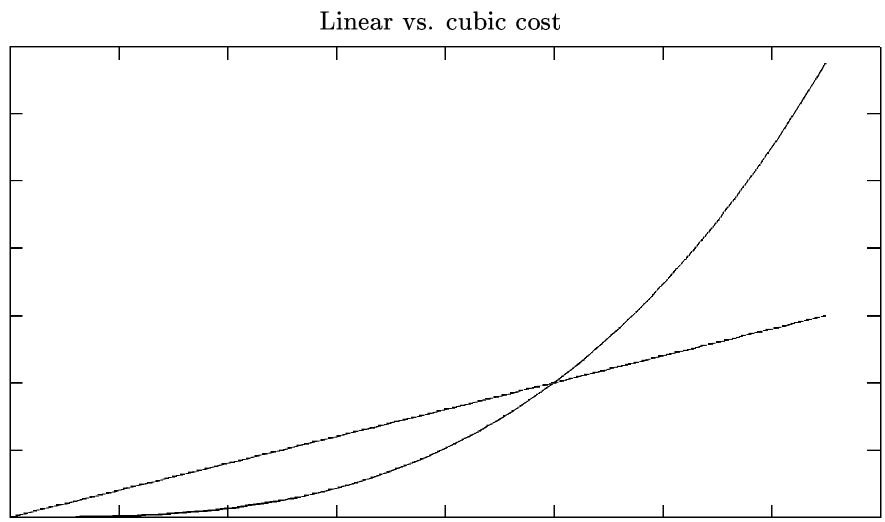
The Purpose of a Modern Numerical Algorithm

- Produce engineering (physical, biochemical, etc.) results with a minimum expenditure of *human* time

- CPU time is irrelevant *as long as it is affordable* (!!!)

- Note to the algorithm designer: torpedoes should not be aimed at the present location of the ship!

Illustration: Algorithms with CPU time estimates $O(n^3)$, $O(n \cdot log(n))$



- To a large extent, the choice of the algorithm is determined by the power of one's computer (!!)

What *do* We Want from a Numerical Algorithm?

- Speed, in the asymptotic sense
- Adaptivity
- Robustness
- Rapid convergence and controlled accuracy: fallacy of the "engineering accuracy" argument; high cost of low precision
- Surprise: adaptivity implies controlled condition numbers; integral vs. differential equations; fast algorithms
- Related surprise: in order to be efficient (or even simply useful), certain algorithms have to be fairly complicated (think about modern cars)

Subject of This Talk

- Talk is cheap - examples are needed
- FMMs for the Helmholtz (Maxwell's) Equation in the "wideband" environment - explain
- Something of a misnomer - *mea culpa*!
- Disclaimer: Boeing, HRL, Illinois, MadMax...
- A post-mortem for a project
- Connections with Moore's law, etc.

FMM for the Helmholtz (Maxwell) Equation

- Function: evaluate potentials, fields, etc. of charge distributions. N^2 vs. N or $N \cdot log(N)$, or $N \cdot (log(N))^2$. . .

- Does not provide discretizations, integral formulations, iterative solvers, etc. (left to the user as an exercise)

- Indifferent to all of these issues - explain

- In reality, consists of two procedures. One is used on the subwavelength scale (or in low-frequency environments), the other is used in the high-frequency environment; transition is seamless

Low-Frequency (Subwavelength) Environment

- Similar to Laplace - explain

- Very simple “bare-bones” scheme, more involved “modern” versions

- Fairly fast: (several times slower than the Laplace FMM) for groups up to 4 λ or so (define the groups)

- Break-even points

- Behavior as groups increase

- Serious deterioration for groups greater than 5 to 8 λ

- Fairly simple implementations produce acceptable results

High-Frequency Environment

- Not at all similar to the Laplace case: "oscillatory behavior"

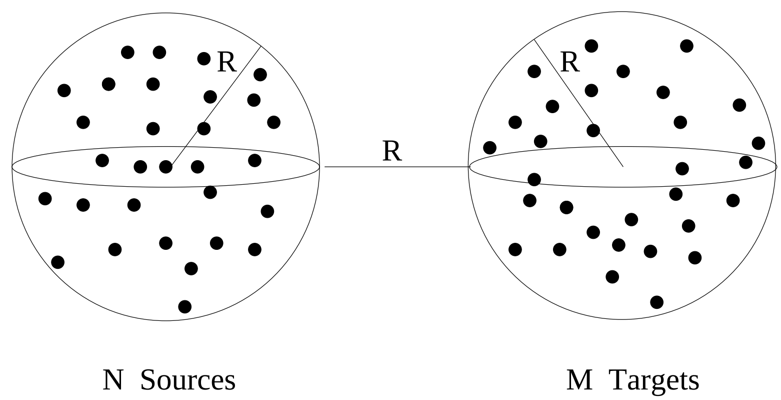
- Example with the Moon

- "At a fixed number of points per λ, the rank of each submatrix is proportional to its size" - not quite true, Michielssen counterexample

- How bad is it?

- Let us see

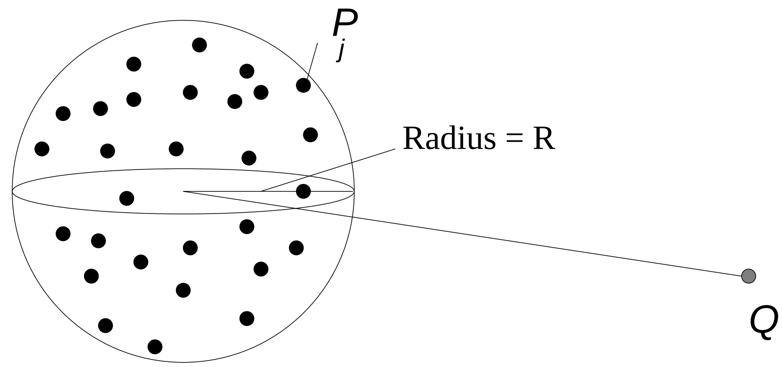
At the Bottom of the Scheme



$$V(Q_i) = \sum_{j=1}^{N} q_j \, \frac{e^{ik\|Q_i - P_j\|}}{\|Q_i - P_j\|}$$

Direct evaluation requires $O(NM)$ work

At the Bottom of the Scheme II



$$V(Q) = V(r,\theta,\phi) \approx \sum_{n=0}^{p} \sum_{m=-n}^{n} M_n^m Y_n^m(\theta,\phi) h_n(kr),$$

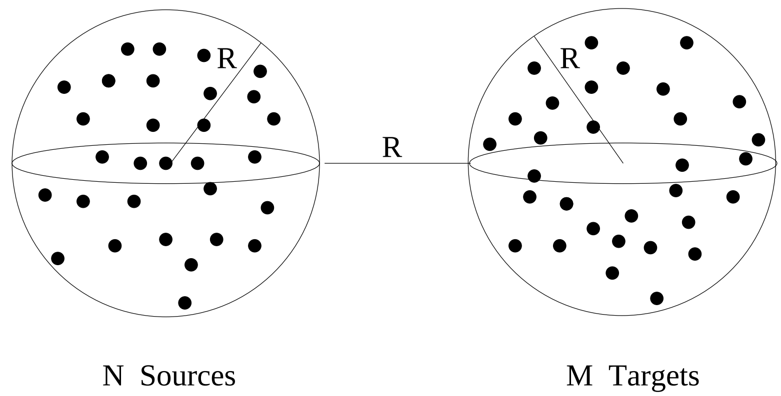
with multipole moments

$$M_n^m = \sum_{j=1}^{N} q_j Y_n^{-m}(\theta_j,\phi_j) j_n(kr), \quad P_j = (r_j,\theta_j,\phi_j)$$

In the low frequency regime, the error in the multipole approximation decays like $(R/|Q|)^{p+1}$.

For our simple example, $R/|Q| < 1/2$, so that setting $p = \log_2(\frac{1}{\varepsilon})$ yields a precision of ε.

At the Bottom III



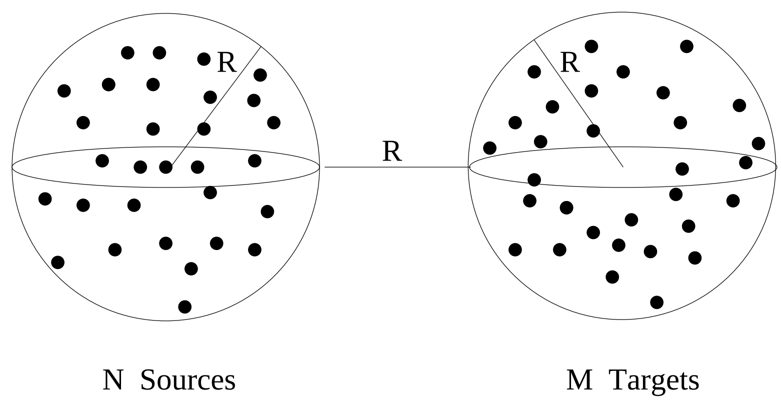
- Evaluate multipole coefficients M_n^m for $n = 0, \ldots, p$

- Evaluate expansion at target points Q_j, for $j = 1, \ldots, M$

- Total operation count: $p^2 \cdot (N + M) = (N + M) \cdot \log^2(\frac{1}{\epsilon})$

- The schemes depend critically on p^2 being much smaller than N

Hard Life at High Frequencies

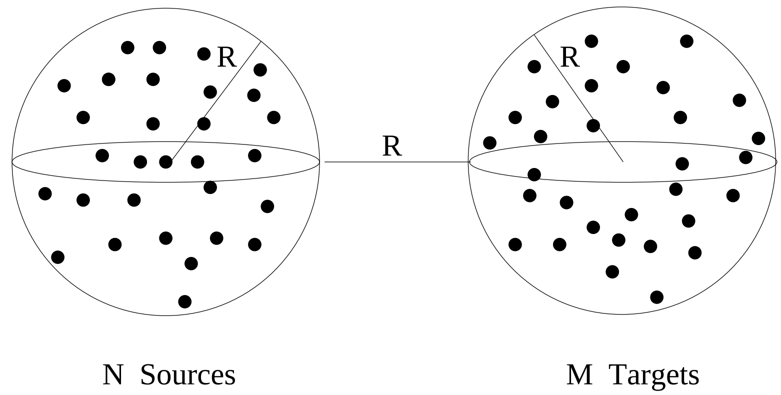


$$V(r,\theta,\phi) \approx \sum_{n=0}^{p} \sum_{m=-n}^{n} M_n^m Y_n^m(\theta,\phi) h_n(kr)$$

- Coefficients M_n^m do not start decaying until $n > |k \cdot R|$, after which decay is extremely rapid

- Condition $p > |k \cdot R| + O(|k \cdot R|^{1/3})$ is needed if we are to have any accuracy at all

Hard Life II



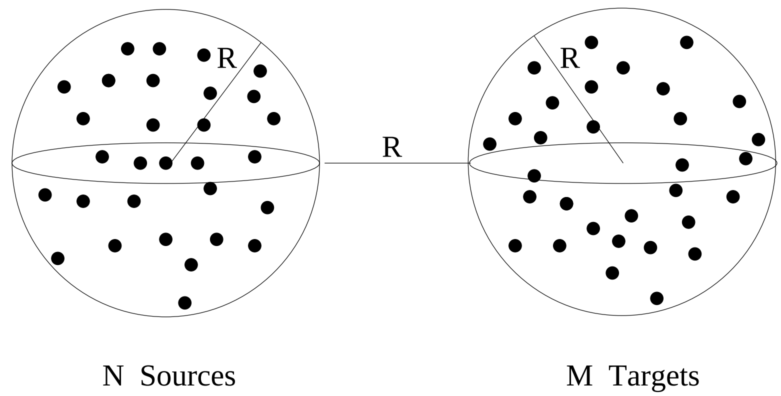
- p is proportional to $\frac{R}{\lambda}$

- In BIE discretizations: fixed number of nodes per λ^2

- Thus, total number of elements in the expansion is of the same order as N

- None of the $O(N \cdot log(N))$ schemes (Barnes-Hut, etc.) will work in this regime

Hard Life III



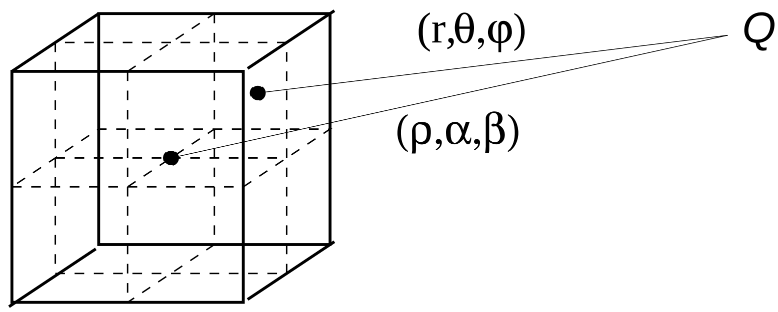
- Another way to put it: the rank approach will not work because the ranks are high

- Cooked goose, vicious gloating

- The situation is a little better when volume distributions and volume integrals are considered, but not enough - and there is FFT-based competition

- What about order N algorithms (FMMs)?

Translation Operators ($h \to h$)



$$\sum_{n=0}^{p} \sum_{m=-n}^{n} M_n^m Y_n^m(\theta, \phi) h_n(kr) \to$$

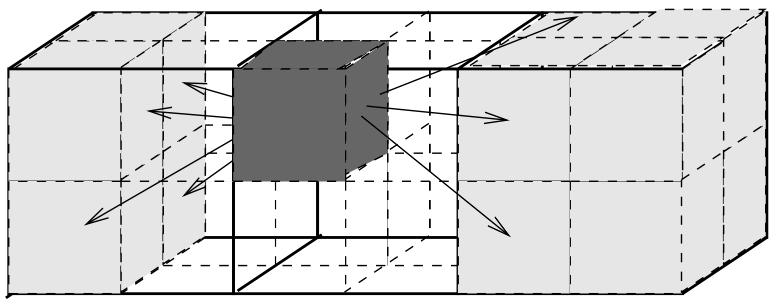
$$\to \sum_{n=0}^{p} \sum_{m=-n}^{n} N_n^m Y_n^m(\alpha, \beta) h_n(k\rho)$$

- Cost: $O(p^4)$

- $O(p^3)$ via "point and shoot" procedure

- Fatal in the BIE environment

Translation Operators ($h \to j$)

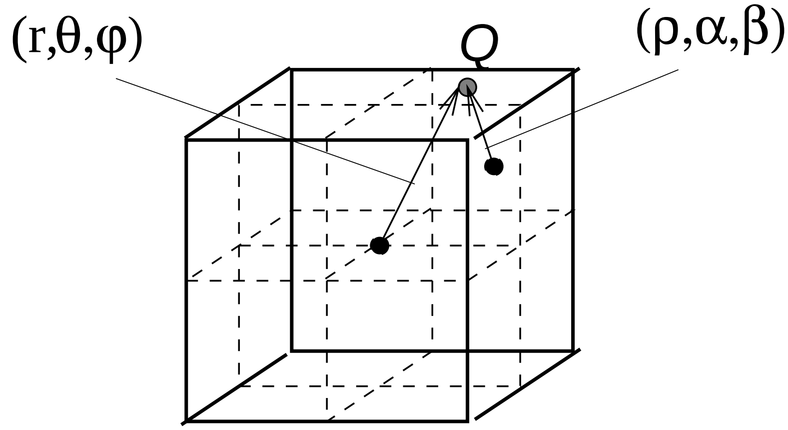


$$\sum_{n=0}^{p} \sum_{m=-n}^{n} M_n^m Y_n^m(\theta,\phi) h_n(kr) \to$$
$$\to \quad \sum_{n=0}^{p} \sum_{m=-n}^{n} L_n^m Y_n^m(\alpha,\beta)\, j_n(k\rho)$$

- No better than $h \to h$

- Dominant type of translation in an FMM

Translation Operators ($j \to j$)



$$\sum_{n=0}^{p} \sum_{m=-n}^{n} L_n^m Y_n^m(\theta,\phi)\, j_n(kr) \to$$

$$\to \sum_{n=0}^{p} \sum_{m=-n}^{n} O_n^m Y_n^m(\alpha,\beta)\, j_n(k\rho)$$

- Same as $h \to h$

A Grim Observation

- Ranks of translation operators in the high-frequency Helmholtz (Maxwell's, etc.) environment are proportional to the sizes of the groups in wavelengths (with subtle exceptions - Michielssen)

- For surface distributions of charges, any FMM that as much as creates translation operators will be of order at least $O(N^2)$ - horror!

- Translation operators in their "point and shoot" form reduce best possible order to $O(N^{3/2})$ - not nearly good enough

- Classical translation operators are of little use in the construction of Helmholtz FMMs, except at low frequencies

What Is Needed

- Bases in which translation operators are diagonal, or at least very sparse
- Transitions between such bases must be very sparse
- Transitions between the standard representations (partial wave expansions) and the new bases must be very sparse
- Alternatively, it should be possible to carry out the whole procedure in the "dual" bases
- Where does one find such paragons?

A Pleasant Observation

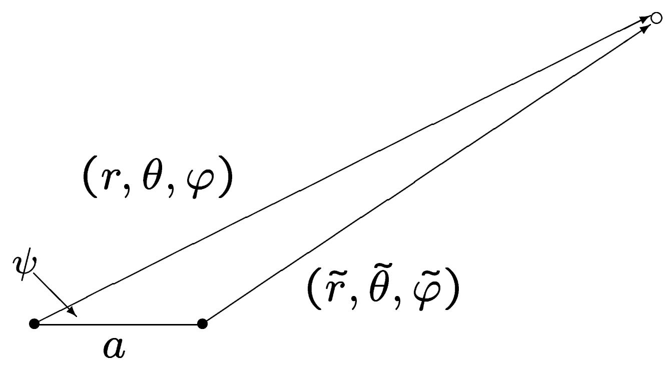
- All translation operators on a given level are diagonalized by the same unitary operator

- All diagonal forms are available analytically

- Transitions between bases (corresponding to different levels) can be done in a “fast” manner

- The whole procedure is quite simple, as long as it is understood in an appropriate weak sense

Radiation Potentials and T_{hh}



$$P(r,\theta,\varphi) = \sum_{n=-\infty}^{+\infty} \sum_{m=-n}^{n} M_n^m Y_n^m(\theta,\phi) h_n(kr)$$

$$P(\tilde{r},\tilde{\theta},\tilde{\varphi}) = \sum_{n=-\infty}^{\infty} \sum_{m=-n}^{n} \tilde{M}_n^m Y_n^m(\tilde{\theta},\tilde{\phi}) h_n(k\tilde{r})$$

Sommerfeld condition:

$$\lim_{r\to\infty} P(r,\theta,\varphi) \cdot r \cdot e^{-i\cdot k\cdot r} = F(\theta,\phi)$$

Observation

The mapping

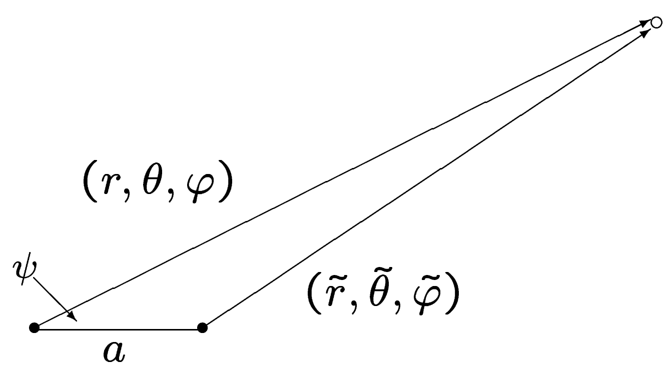
$$U : \{M_n^m\} \to F(\theta, \phi)$$

diagonalizes the translation operator

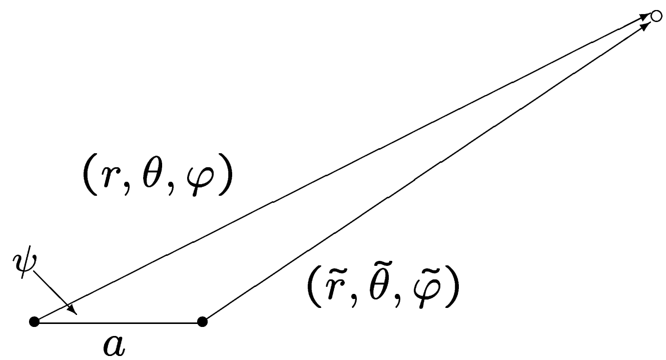
$$T_{hh} : \{M_n^m\} \to \{\tilde{M}_n^m\}$$

On the diagonal

$$e^{i \cdot k \cdot a \cdot cos(\psi)}$$



Proof:



For large r,

$$(\tilde{\theta}, \tilde{\varphi}) \sim (\theta, \varphi),$$

which means that the mapping

$$U^{-1} \circ T_{hh} \circ U : F \to \tilde{F}$$

is diagonal. For large r,

$$\tilde{r} - r \sim a \cdot cos(\psi),$$

and

$$\left(U^{-1} \circ T_{hh} \circ U\right)(\theta, \varphi) = e^{i \cdot k \cdot a \cdot cos(\psi)}$$

What *Is* U?

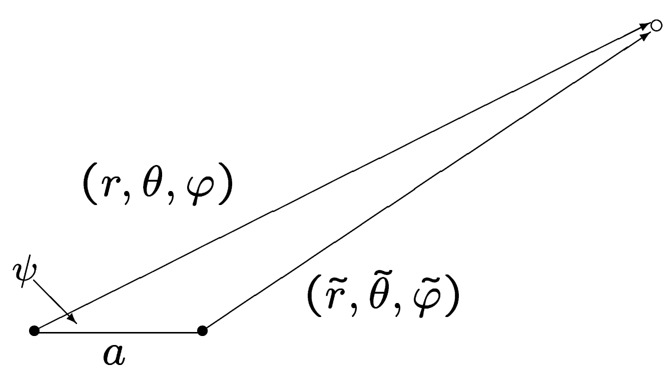
For large r

$$h_m(kr) \sim \frac{e^{i \cdot k \cdot r}}{k \cdot r}$$

(up to some powers of i), and

$$\sum_{n=0}^{p} \sum_{m=-n}^{n} M_n^m Y_n^m(\theta, \varphi) h_n(kr) \sim$$

$$\sim \quad \frac{e^{i \cdot k \cdot r}}{k \cdot r} \sum_{n=0}^{p} \sum_{m=-n}^{n} M_n^m Y_n^m(\theta, \varphi) = F(\theta, \varphi)$$



What Have We Achieved?

- T_{hh} is a spherical convolution; it is diagonalized by the spherical harmonic transform; its diagonal form is a function living on S^2.

- T_{hh} is unitary; its diagonal is $e^{i \cdot k \cdot a \cdot cos(\psi)}$

- Direct result of the Sommerfeld condition, and has been known for a long time

- And what about T_{jj} and T_{hj}?

Diagonalizing T_{jj}

For large r

$$j_m(kr) \sim \frac{cos(k \cdot r)}{k \cdot r}$$

(up to some phase corrections), and

$$\sum_{n=0}^{p} \sum_{m=-n}^{n} M_n^m Y_n^m(\theta, \varphi) j_n(kr) \sim$$

$$\sim \quad \frac{cos(k \cdot r)}{k \cdot r} \sum_{n=0}^{p} \sum_{m=-n}^{n} M_n^m Y_n^m(\theta, \varphi) = F(\theta, \varphi)$$

- A Sommerfeld condition of sorts

Diagonalizing T_{jj} II

$$\sum_{n=0}^{p} \sum_{m=-n}^{n} M_n^m Y_n^m(\theta,\varphi) j_n(kr) \sim$$

$$\sim \quad \frac{cos(k \cdot r)}{k \cdot r} \sum_{n=0}^{p} \sum_{m=-n}^{n} M_n^m Y_n^m(\theta,\varphi)$$

- Makes no physical sense whatsoever

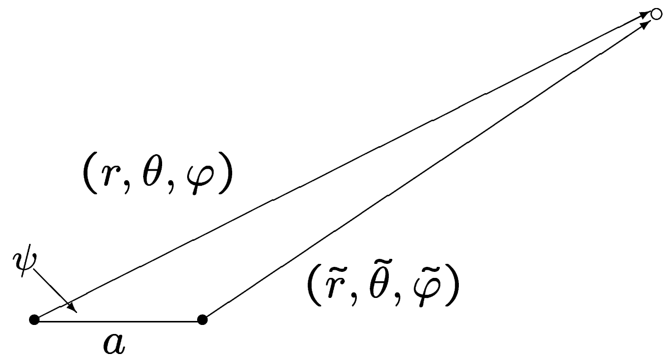
- As $p \to \infty$, the limit usually does not even exist!

- First truncate, then take the limit; for this, we will pay later

- Diagonalized by the harmonic transform, same as T_{hh}; the same $e^{i \cdot k \cdot a \cdot cos(\psi)}$ on the diagonal

- Purely formal expedient

Corollary



Far-field signature of a unit charge is given by the formula

$$F(\theta,\varphi) = e^{i\cdot k\cdot a\cdot cos(\psi)};$$

The potential at the point (a,θ,φ) of the $J-expansion$ with the far-field signature σ is given by the formula

$$P(a,\theta,\varphi) = \int_{S^2} \sigma(\tilde{\theta},\tilde{\varphi}) \cdot e^{-i\cdot k\cdot a\cdot cos(\psi)} ds$$

What about T_{hj}?

$$
\begin{tikzcd}
\{M_n^m\} \arrow[rr, "\tilde{T}_{hj}"] \arrow[d, "T_{hh}"'] \arrow[drr, "T_{hj}"] & & \{\tilde{N}_n^m\} \arrow[d, "T_{jj}"] \\
\{\tilde{M}_n^m\} \arrow[rr, "\tilde{T}_{hj}"'] & & \{N_n^m\}
\end{tikzcd}
$$

- Operators T_{hh}, T_{jj} are diagonal in the far-field representation, and $T_{hh} = T_{jj}$

- Furthermore,

$$T_{jj} \circ \tilde{T}_{hj} = \tilde{T}_{hj} \circ T_{hh}$$

- Inevitable consequences

- Commutative diagrams, morality, etc.

What Is On The Diagonal?

$$\sum_{n=0}^{\infty} (2\,n+1)\,h_n(k\,\rho)\,P_n(\cos(\psi))$$

- "Addition theorem"

- Abramovitz and Stegun

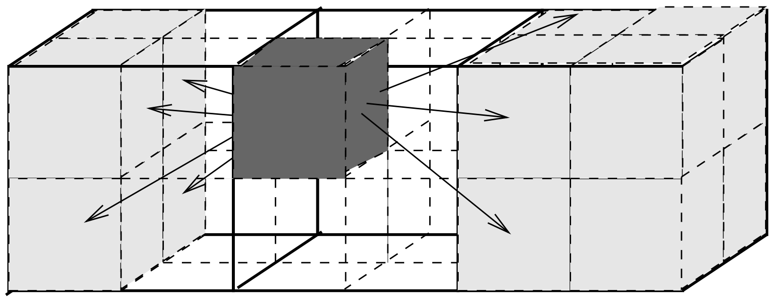
- Series above is divergent; truncation, accuracy, dynamic range, etc.

- Usual situation with convolutions with divergent sequences

- Analysis is a little detailed; results are summarized below

- Variations: beam-like translation operators, etc.

Summary



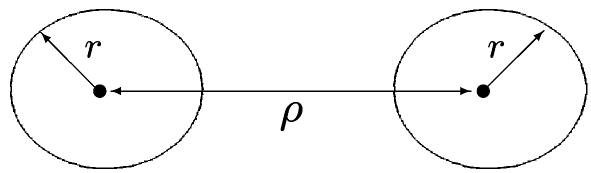
- All translations within one level are diagonalized by the far-field signature

- Far-field signatures of charge (dipole, whatever) distributions are given by simple formulae, and fairly inexpensive to evaluate

- Far-field signatures are smooth functions on the sphere, and can be represented by tables of their values - elaborate

- Transitions between levels involve interpolation and filtering of functions on the sphere. Interpolation is easy; filtering has been taken care of (Alpert-Jacob-Chien Algorithm, Dembart and VR, etc.)

"Low-Frequency Break-Down"



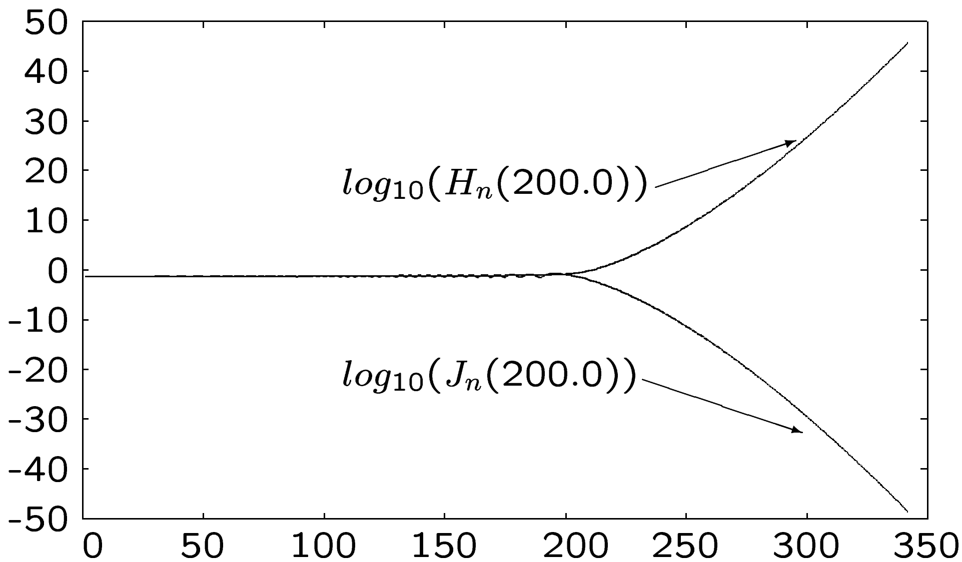
- Outgoing h-expansion behaves as $j_n(k\,r)$

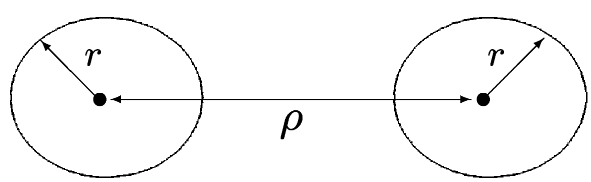
- Incoming j-expansion is a convolution of the outgoing h-expansion with the original (physical space) translation operator; the latter behaves as $h_n(k\,\rho)$

- The potential at a point within the target sphere (circle) is obtained as an inner product of the incoming j-expansion with a sequence behaving as $j_n(k\,r)$

"Low-Frequency Break-Down" II

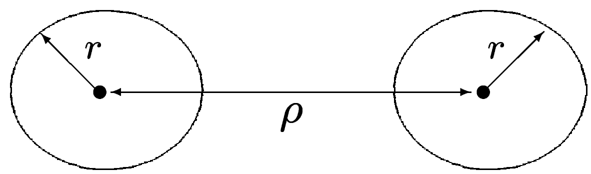
Behavior of Bessel Functions:





- When convolutions are done explicitly, the procedure is numerically stable as long as the spheres do not intersect (physics never lies, even if it takes a conspiracy)

“Low-Frequency Break-Down” III



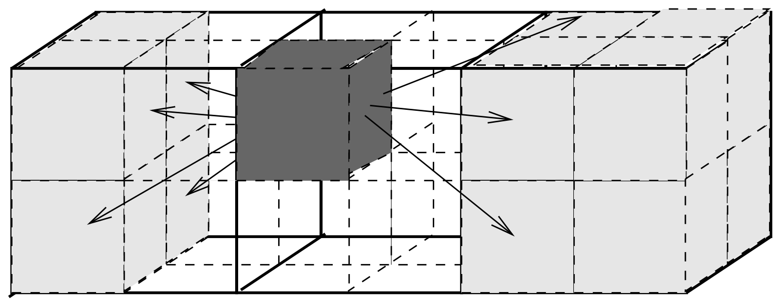
- When convolutions are done via Fourier Transforms (or via spherical transforms) the *dynamic range* of each sequence must not be large. In other words, $J_n(k\,r)$ must **implode** before $H_{2n}(k\,\rho)$ **explodes**

- For sufficiently large $k\,r$, the condition $\rho \geq 3\,r$ is sufficient. For smaller r, greater separation is needed

- Separation depends on the required accuracy, $k\,r$, and the machine ε - explain

- In this case, a table is worth a thousand theories

"Low-Frequency Break-Down": Table



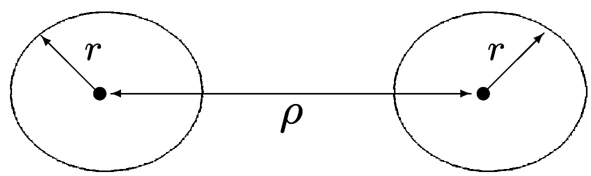
- Double precision calculations

  3 digits $0.25\,\lambda$ side of the cube
  6 digits $3.50\,\lambda$ side of the cube
  9 digits $12.0\,\lambda$ side of the cube

- Similarity with evaluation $sin(a \cdot x)$
  via Taylor series - explain

- Marginal improvements are possible

“Low-Frequency Break-Down”: Remedy



- What does one do in the subwavelength regime?

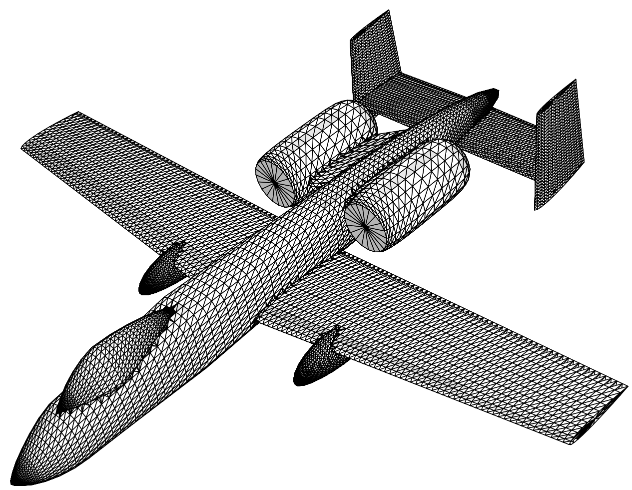
- Use the low-frequency version of the FMM

- Transition to the high-frequency (diagonal) version at the appropriate point

- We have not tried to play with the size of the buffer

Numerical Examples

A-10



- 50 wavelengths in size
- Smallest triangle: 1.06E-6 λ
- Largest triangle: 2.86E-1 λ
- Number of triangles: $706,300$
- Single node per triangle

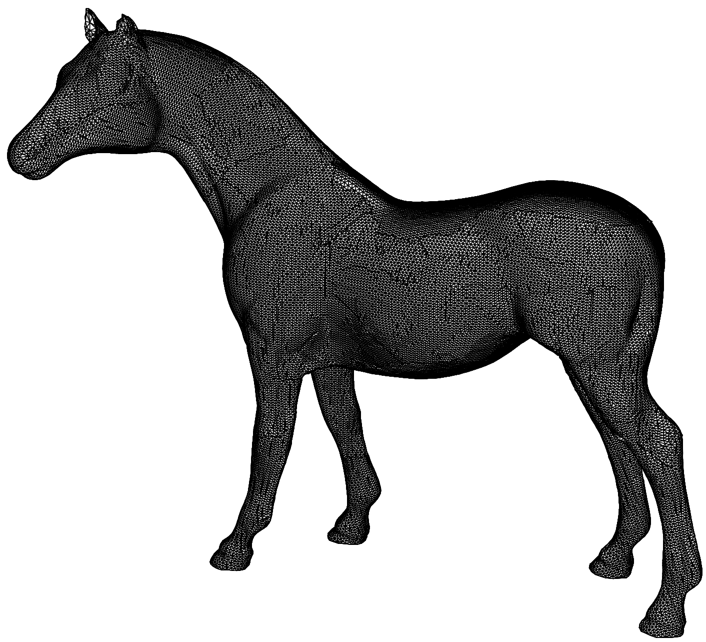
A-10 - Helmholtz

T (dir.)	Acc.	Error (pot.)	Error (grad.)	T (sec.)	Mem. (Mb)
337329	10^{-3}	0.43E-3	0.56E-3	485	300
337329	10^{-6}	0.48E-6	0.50E-6	1291	790
337329	10^{-9}	0.11E-9	0.95E-10	2947	1143

A-10 - Laplace

T (dir.)	Acc.	Error (pot.)	Error (grad.)	T (sec.)	Mem. (Mb)
60590	10^{-3}	0.27E-3	0.37E-4	48.3	211
60590	10^{-6}	0.19E-6	0.43E-7	119	292
60590	10^{-9}	0.85E-10	0.61E-11	2437	376

Horse



- 50 wavelengths in size
- Smallest triangle: 9.34E-3 λ
- Largest triangle: 3.27E-1 λ
- Number of triangles: $872,694$
- Single node per triangle

Horse - Helmholtz

T (dir.)	Acc.	Error (pot.)	Error (grad.)	T (sec.)	Mem. (Mb)
646143	10^{-3}	0.65E-3	0.31E-3	672	549
646143	10^{-6}	0.66E-6	0.92E-7	1832	1111
646143	10^{-9}	0.33E-9	0.33E-11	3515	2027

Horse - Laplace

T (dir.)	Acc.	Error (pot.)	Error (grad.)	T (sec.)	Mem. (Mb)
107833	10^{-3}	0.91E-3	0.57E-3	63.7	328
107833	10^{-6}	0.46E-6	0.31E-6	139.7	322
107833	10^{-9}	0.25E-9	0.10E-9	298	584

Sphere

- 50 wavelengths in size
- Smallest triangle: 4.91E-2 λ
- Largest triangle: 6.27E-2 λ
- Number of triangles: $619,520$
- Single node per triangle

Sphere - Helmholtz

T (dir.)	Acc.	Error (pot.)	Error (grad.)	T (sec.)	Mem. (Mb)
324381	10^{-3}	0.27E-3	0.19E-3	521	416
324381	10^{-6}	0.15E-6	0.42E-7	1358	914
324381	10^{-9}	0.91E-10	0.24E-10	2873	1474

Sphere - Laplace

T (dir.)	Acc.	Error (pot.)	Error (grad.)	T (sec.)	Mem. (Mb)
52936	10^{-3}	0.79E-3	0.90E-3	45	245
52936	10^{-6}	0.33E-6	0.45E-6	97.7	244
52936	10^{-9}	0.19E-9	0.12E-9	223	402

Cube

- 50 wavelengths in size
- Smallest triangle: 9.12E-2 λ
- Largest triangle: 9.12E-2 λ
- Number of triangles: $668,352$
- Single node per triangle

Cube - Helmholtz

T (dir.)	Acc.	Error (pot.)	Error (grad.)	T (sec.)	Mem. (Mb)
376950	10^{-3}	0.97E-3	0.74E-3	393	364
376950	10^{-6}	0.73E-6	0.26E-7	1022	1295
376950	10^{-9}	0.23E-9	0.17E-10	2077	1001

Cube - Laplace

T (dir.)	Acc.	Error (pot.)	Error (grad.)	T (sec.)	Mem. (Mb)
56433	10^{-3}	0.94-3	0.60E-3	52	201
56433	10^{-6}	0.41E-6	0.34E-6	132	272
56433	10^{-9}	0.28E-9	0.17E-9	231	362

Observations

- A fairly mature technology

- Unlike the Laplace case, it is technical (as opposed to incantational), even on the most basic level - explain

- It is not enough to "invent" an order n (or $n \cdot log(n)$, or whatever) scheme any more - constants matter

- Robustness and ease of use, accuracy control, careful testing, implementation practices, etc.

- A little mathematics goes a long way - implications

- Algorithms are becoming technical and involved; have to be developed by competent groups

- An engineering discipline vs. black art

Conclusions

- Fast BIE solvers for Elliptic, parabolic, hyperbolic equations

- "Fast" algorithms for fast computers

- A different collection of collateral issues: surface descriptions, high-order discretizations, volume integrals, etc.

- Other environments involving "fastness" - Moore's law and its consequences

- There are still some freebies left!

- What else?

Conclusions II

- Direct vs. iterative solvers

- BIEs in two dimensions

- Direct solvers for the Lippman-Schwinger equations, non-oscillatory and otherwise - scope and promise

- “Fast” SVDs and eigendecompositions

- DIRECT SOLVERS!!!

- Applications: Singular perturbation problems, problems in the vicinity of resonances, non-linear problems, inverse scattering

- INVERSE SCATTERING!!!