

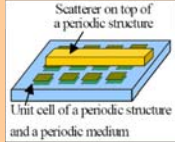
Fast Evaluation of Optical Fields in Complex Nanophotonic Structures

Shaojing Li, Boris Livshitz, Derek Van Orden, Yeshaiahu Fainman, and Vitaliy Lomakin

Department of Electrical and Computer Engineering, University of California, San Diego

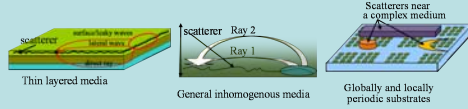
Motivation: Complex nanophotonics

- Computational modeling is essential for fabrication, characterization and optimization of photonic systems
- Many nanophotonic structures and systems have complex geometries and materials
- New computational methods be developed with orders of magnitude improvement in computation speed



Hybrid methods for Green's Functions (1)

- Develop representation for a set of canonical problems
 - Layered medium Green's functions
 - Green's functions in general inhomogeneous media
 - Periodic Green's functions for periodic and aperiodic excitations
 - Etc.
- Importance
 - Study of wave propagation
 - Use with convolution accelerators for fast IE methods



Non-Uniform Grid Method (NGM)

- NG algorithms evaluate potentials due to known sources in $O(N \log N)$ or $O(N)$ operations*

- NG algorithms are based on field compensation, sampling, and interpolations

- Advantages of NG algorithms:

- highly adaptive to problem geometry
- flexible in error control
- applied to mixed-frequency problems
- Can be extended to very complex configurations once "good" representations of Green's functions are obtained
- NG schemes exist for
 - free space* (statics, frequency domain & time domain dynamics fields)
 - layered media/substrates are under development
 - More complicated configurations are to be developed



*Boag; Boag & Livshitz; Boag, Lomakin, & Michielssen

Parallelization with GPUs

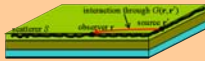
- Classic hardware in scientific computing:
 - Single-core CPU (saturates after Pentium D time)
 - Multi-core CPU (mainstream, 2-core/4-core)
 - Cluster from many cores (powerful, but expensive and power hungry)
- Alternatives for general purpose scientific computing:
 - General Purpose Graphic Processing Unit (GPGPU)
 - nVidia (GeForce, Quadro, Tesla series GPU, supported by CUDA)
 - ATI (Radeon series GPU, supported by Stream technologies)
 - Cell processor from IBM (complicated software environment)



Motivation: Integral equation (IE) methods (1)

- Integral equation (IE) methods are well suited to analyze fields in complex nanophotonic structures
- An integral equation for a surface problems written symbolically

$$\int_S G(\mathbf{r}, \mathbf{r}') J(\mathbf{r}') ds' = \psi(\mathbf{r}); \quad \mathbf{r} \in S$$



- Information anytime, synchronizing their behavior and helping each other
- Shared memory: fastest memory on board but also limited in amount. This level of memory is as fast as register and can be accessed by a block of threads at the same time.
- Coalesced accessing of global memory: global memory has "narrower" bandwidth (140 Gbytes/sec)
- High accuracy and speed
- Parallelization techniques to accelerate computations
 - Are crucial to allow for continued scaling, as the frequency of each core is saturating
 - Explore novel architectures including Central Processing Units, Graphics Processing Units, and their heterogeneous system combinations

Hybrid methods for Green's Functions (2)

- The Green's function is given in terms of a few wave species with known asymptotic behavior
 - Eigen modes: Surface and leaky waves, whispering gallery modes
 - Rays: Conventional ray theory, geometric theory of diffraction, uniform theory of diffraction transition functions
 - Beams: Gaussian beams, higher order beams, time domain beams
- Benefits
 - Optimal representations take into account the physical behavior of
 - Avoid bank conflict when operating shared memory
 - Uniform distribution of computational burden among threads
 - Avoiding leaving complicated judgment job and highly divergent branching

Basics of Non-Uniform Grid Method

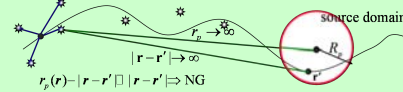
- Ideas behind the NGM

- Computational domain is partitioned into subdomains. Interactions between the subdomains is through sparse grids and interpolation
- Non-uniform grid construction

$$\psi_p = \int_{S_p} d\mathbf{r}' \frac{e^{-jk_p |\mathbf{r} - \mathbf{r}'|}}{4\pi |\mathbf{r} - \mathbf{r}'|} \sigma(\mathbf{r}', t) \Rightarrow \tilde{\psi}_p(\mathbf{r}) = \int_{S_p} \sigma(\mathbf{r}') \frac{e^{-jk_p |\mathbf{r} - \mathbf{r}'|}}{|\mathbf{r} - \mathbf{r}'|} ds'$$

$$|\mathbf{r} - \mathbf{r}'| \rightarrow \tilde{\psi}_p \text{ varies slowly} \Rightarrow \text{sparse grid and interpolations}$$

$$\Rightarrow \text{interpolation with } \Delta \phi_p = (\Omega_p k_p R_p)^{-1}, \Delta(r_p) = (\Omega_p k_p R_p)^{-1} \Rightarrow \psi_p(\mathbf{r}) = \tilde{\psi}_p(\mathbf{r}) \frac{e^{-jk_p |\mathbf{r} - \mathbf{r}'|}}{r_p(\mathbf{r})}$$



Parallelization with GPUs: nVidia cards

- Cheap in price
 - ~\$250 GeForce GTX 280 vs. ~\$1500 Intel Xeon X5482
- Easy to program
 - nVidia CUDA development platform, extension to C language
- Multi-core system
 - 240 stream processors in GTX 280 group into 30 multi-processors that can run a batch of threads.
- Memory Architecture
 - 16Kbytes shared memory for a group of threads to cooperate, may be as fast as registers in CPUs.



Motivation: Integral equation methods (2)

- Properties of IEs

- Information anytime, synchronizing their behavior and helping each other
- Shared memory: fastest memory on board but also limited in amount. This level of memory is as fast as register and can be accessed by a block of threads at the same time.
- Coalesced accessing of global memory: global memory has "narrower" bandwidth (140 Gbytes/sec)
- High accuracy and speed
- Parallelization techniques to accelerate computations
 - Are crucial to allow for continued scaling, as the frequency of each core is saturating
 - Explore novel architectures including Central Processing Units, Graphics Processing Units, and their heterogeneous system combinations

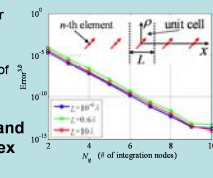
Green's Functions: Periodic arrays

- Periodic free space Green's function representation

$$G = \sum_{n=-\infty}^{\infty} \frac{e^{-jR_n \sqrt{k_0^2 - k_x^2}}}{4\pi R_n} \int_{-\infty}^{\infty} \frac{e^{-j(k_x x + k_y y + k_z z)}}{2Lj} \frac{1}{\sqrt{1 - e^{-j(k_x^2 + k_y^2)}}} e^{-jR_n \sqrt{k_0^2 - k_x^2}} \frac{e^{-jR_n \sqrt{k_0^2 - k_x^2}}}{4\pi k_z} k_y dk_y$$

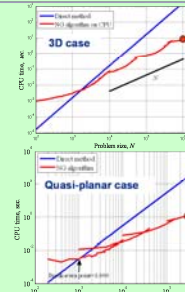
- Results

- Rapid convergence (e.g., ~7 terms for are needed for single precision)
- Small number of wave species
- Clear phase and amplitude behavior of the wave species
- Has been extended to periodic structures near layered media, and will be extended to more complex configurations.



Results for static NGM

- Complexity and memory storage scale linearly
- Spatially adaptive, e.g. faster for quasi-planar structure
- Speed up highlights:
 - Quasi-planar problem with 300,000 sources requires only 1 sec! (2.5 sec for 3D)
 - 3D problem with one million sources requires only 8 sec! (~30 min for DM)



All presented data are obtained on Dell Xeon 2.66GHz Quad Core Workstation

Parallelization with GPUs: Programming

- Programming Techniques:
 - Threads and thread block: a block of threads can share information anytime, synchronizing their behavior and helping each other.
 - Shared memory: fastest memory on board but also limited in amount. This level of memory is as fast as register and can be accessed by a block of threads at the same time.
 - Coalesced accessing of global memory: global memory has "narrower" bandwidth (140 Gbytes/sec)
 - Avoid bank conflict when operating shared memory
 - Uniform distribution of computational burden among threads
 - Avoiding leaving complicated judgment job and highly divergent branching

Motivation: Critical components for IEs

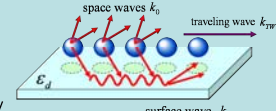
- Using IEs to analyze large-scale complex nanophotonic structures requires three important components:

- Efficient representations for Green's Functions: information anytime, synchronizing their behavior and helping each other.
- Shared memory: fastest memory on board but also limited in amount. This level of memory is as fast as register and can be accessed by a block of threads at the same time.
- Coalesced accessing of global memory: global memory has "narrower" bandwidth (140 Gbytes/sec)
- High accuracy and speed
- Parallelization techniques to accelerate computations
 - Are crucial to allow for continued scaling, as the frequency of each core is saturating
 - Explore novel architectures including Central Processing Units, Graphics Processing Units, and their heterogeneous system combinations

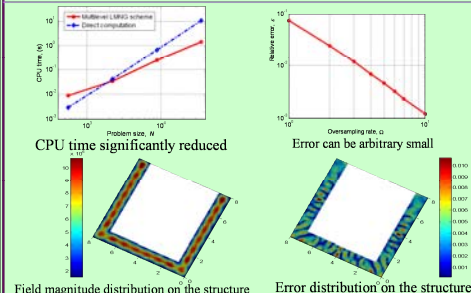
Green's Functions: Study of wave phenomena

- Green's functions has been used to study propagation and radiation from linear chains of nanoparticles near a dielectric slab
- Physical phenomena identified

- The array supports traveling waves that may remain bound to the array or leak into surface or space waves.
- Highly tunable
- Possible applications to surface wave microscopy



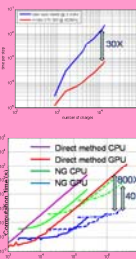
Results for layered medium dynamic NGM



Field magnitude distribution on the structure Error distribution on the structure

Parallelization with GPUs: Results for statics

- GPU-accelerated direct matrix-vector multiplication
 - Easy to implement
 - ~30x speed up
- GPU-accelerated NGM for static field computation
 - Requires rethinking the NGM code
 - Up to ~80x speed-up
 - Up to ~100 memory savings
 - E.g. $N=3e6$ at 0.8 sec at 1e-3 RMS error
 - Simple desktop = a cluster!



All presented data is obtained on Intel Xeon 3.2GHz CPU and nVIDIA GTX 280 GPU