

FAST CALCULATION OF THE LOMB-SCARGLE PERIODOGRAM USING GRAPHICS PROCESSING UNITS

R. H. D. TOWNSEND

Department of Astronomy, University of Wisconsin-Madison, Sterling Hall, 475 N. Charter Street, Madison, WI 53706, USA;
 townsend@astro.wisc.edu

Draft version October 19, 2010

ABSTRACT

I introduce a new code for fast calculation of the Lomb-Scargle periodogram, that leverages the computing power of graphics processing units (GPUs). After establishing a background to the newly emergent field of GPU computing, I discuss the code design and narrate key parts of its source. Benchmarking calculations indicate no significant differences in accuracy compared to an equivalent CPU-based code. However, the differences in performance are pronounced; running on a low-end GPU, the code can match 8 CPU cores, and on a high-end GPU it is faster by a factor approaching thirty. Applications of the code include analysis of long photometric time series obtained by ongoing satellite missions and upcoming ground-based monitoring facilities; and Monte-Carlo simulation of periodogram statistical properties.

Subject headings: methods: data analysis — methods: numerical — techniques: photometric — stars: oscillations

1. INTRODUCTION

Astronomical time-series observations are often characterized by uneven temporal sampling (e.g., due to transformation to the heliocentric frame) and/or non-uniform coverage (e.g., from day/night cycles, or radiation belt passages). This complicates the search for periodic signals, as a fast Fourier transform (FFT) algorithm cannot be employed. A variety of alternatives have been put forward, the most oft-used being the eponymous Lomb-Scargle (L-S) periodogram developed by Lomb (1976) and Scargle (1982). At the time of writing, NASA’s Astrophysics Data System (ADS) lists 735 and 1,810 publications (respectively) that cite these two papers, highlighting how important the L-S periodogram has proven for the analysis of time series. Recent applications include the search for a link between solar rotation and nuclear decay rates (Sturrock et al. 2010); the study of pulsar timing noise (Lyne et al. 2010); the characterization of quasi-periodic oscillations in blazars (Rani et al. 2010); and the measurement of rotation periods in exoplanet host stars (Simpson et al. 2010).

Unfortunately, a drawback of the L-S periodogram is a computational cost scaling as $\mathcal{O}(N_t^2)$, where N_t is the number of measurements in the time series; this contrasts with the far-more-efficient $\mathcal{O}(N_t \log_2 N_t)$ scaling of the FFT algorithm popularized by Cooley & Tukey (1965). One approach to reducing this cost has been proposed by Press & Rybicki (1989), based on constructing a uniformly sampled approximation to the observations via ‘extirpolation’ and then evaluating its FFT. The present paper introduces a different approach, not through algorithmic development but rather by leveraging the computing power of graphics processing units (GPUs) — the specialized hardware at the heart of the display subsystem in personal computers and workstations. Modern GPUs typically comprise a number of identical programmable processors, and in recent years there has been significant interest in applying these parallel-computing resources to problems across a breadth of scientific disciplines. In the following section, I give a brief history

of the newly emergent field of GPU computing; then, Section 3 reviews the formalism defining the L-S periodogram, and Section 4 presents a GPU-based code implementing this formalism. Benchmarking calculations to evaluate the accuracy and performance of the code are presented in Section 5. The findings and future outlook are then discussed in Section 6.

2. BACKGROUND TO GPU COMPUTING

2.1. Pre-2006: Initial Forays

The past decade has seen remarkable increases in the ability of computers to render complex 3-dimensional scenes at movie frame-rates. These gains have been achieved by progressively shifting the graphics pipeline — the algorithmic sequence of steps that converts a scene description into an image — from the CPU to dedicated hardware within the GPU. To address the inflexibility that can accompany such hardware acceleration, GPU vendors introduced so-called *programmable shaders*, processing units that apply a simple sequence of transformations to input elements such as image pixels and mesh vertices. NVIDIA Corporation were the first to implement programmable shader functionality, with their GeForce 3 series of GPUs (released March 2001) offering one vertex shader and four (parallel) pixel shaders. The release in the following year of ATI Corporation’s R300 series brought not only an increase in the number of shaders (up to 4 vertex and 8 pixel), but also capabilities such as floating-point arithmetic and looping constructs that laid the foundations for what ultimately would become GPU computing.

Shaders are programmed using a variety of specialized languages, such as the OpenGL Shading Language (GLSL; e.g., Rost 2006) and Microsoft’s High-Level Shading Language (HLSL). The designs of these languages are strongly tied to their graphics-related purpose, and thus early attempts at GPU computing using programmable shaders had to map each calculation into a sequence of equivalent graphical operations (see, e.g., Owens et al. 2005, and references therein). In an effort to

overcome this awkward aspect, Buck et al. (2004) developed BrookGPU — a compiler and run-time implementation of the Brook stream programming language for GPU platforms. With BrookGPU, the computational resources of shaders are accessed through a *stream processing* paradigm: a well-defined series of operations (the *kernel*) are applied to each element in a typically-large homogeneous sequence of data (the *stream*).

2.2. Post-2006: Modern Era

GPU computing entered its modern era in 2006, with the release of NVIDIA’s *Compute Unified Device Architecture* (CUDA) — a framework for defining and managing GPU computations without the need to map them into graphical operations. CUDA-enabled devices (see Appendix A of NVIDIA 2010) are distinguished by their general-purpose unified shaders, which replace the function-specific shaders (pixel, vertex, etc.) present in earlier GPUs. These shaders are programmed using an extension to the C language, which follows the same stream-processing paradigm pioneered by BrookGPU. Since the launch of CUDA, other vendors have been quick to develop their own GPU computing offerings, most notably Advanced Micro Devices (AMD) with their *Stream* framework, and Microsoft with their *DirectCompute* interface.

Abstracting away the graphical roots of GPUs has made them accessible to a very broad audience, and GPU-based computations are now being undertaken in fields as diverse as molecular biology, medical imaging, geophysics, fluid dynamics, economics and cryptography (see Pharr 2005; Nguyen 2007). Within astronomy and astrophysics, recent applications include N -body simulations (Belleman et al. 2008), real-time radio correlation (Wayth et al. 2009), gravitational lensing (Thompson et al. 2010), adaptive-mesh hydrodynamics (Schive et al. 2010) and cosmological reionization (Aubert & Teyssier 2010).

3. THE LOMB-SCARGLE PERIODOGRAM

This section reviews the formalism defining the Lomb-Scargle periodogram. For a time series comprising N_t measurements $X_j \equiv X(t_j)$ sampled at times t_j ($j = 1, \dots, N_t$), *assumed throughout to have been scaled and shifted such that its mean is zero and its variance is unity*, the normalized L-S periodogram at frequency f is

$$P_n(f) = \frac{1}{2} \left\{ \frac{\left[\sum_j X_j \cos \omega(t_j - \tau) \right]^2}{\sum_j \cos^2 \omega(t_j - \tau)} + \frac{\left[\sum_j X_j \sin \omega(t_j - \tau) \right]^2}{\sum_j \sin^2 \omega(t_j - \tau)} \right\}. \quad (1)$$

Here and throughout, $\omega \equiv 2\pi f$ is the angular frequency and all summations run from $j = 1$ to $j = N_t$. The frequency-dependent time offset τ is evaluated at each ω via

$$\tan 2\omega\tau = \frac{\sum_j \sin 2\omega t_j}{\sum_j \cos 2\omega t_j}. \quad (2)$$

As discussed by Schwarzenberg-Czerny (1998), P_n in the case of a pure *Gaussian*-noise time series is drawn from

a beta distribution. For a periodogram comprising N_f frequencies¹, the *false-alarm probability* (FAP) — that some observed peak occurs due to chance fluctuations — is

$$Q = 1 - \left[1 - \left(1 - \frac{2P_n}{N_t} \right)^{(N_t-3)/2} \right]^{N_f}. \quad (3)$$

Equations (1) and (2) can be written schematically as

$$P_n(f) = \sum_j \mathcal{G}[f, (t_j, X_j)], \quad (4)$$

where \mathcal{G} is some function. In the classification scheme introduced by Barsdell et al. (2010), this follows the form of an *interact algorithm*. Generally speaking, such algorithms are well-suited to GPU implementation, since they are able to achieve a high arithmetic intensity. However, a straightforward implementation of equations (1) and (2) involves two complete runs through the time series to calculate a single $P_n(f)$, which is wasteful of memory bandwidth and requires $N_f(4N_t + 1)$ costly trigonometric function evaluations for the full periodogram. Press et al. (1992) address this inefficiency by calculating the trig functions from recursion relations, but this approach is difficult to map onto stream processing concepts, and moreover becomes inaccurate in the limit of large N_f . An alternative strategy, which avoids these difficulties while still offering improved performance, comes from refactoring the equations as

$$P_n(f) = \frac{1}{2} \left[\frac{(c_\tau XC + s_\tau XS)^2}{c_\tau^2 CC + 2c_\tau s_\tau CS + s_\tau^2 SS} + \frac{(c_\tau XS - s_\tau XC)^2}{c_\tau^2 SS - 2c_\tau s_\tau CS + s_\tau^2 CC} \right], \quad (5)$$

and

$$\tan 2\omega\tau = \frac{2CS}{CC - SS}. \quad (6)$$

Here,

$$c_\tau = \cos \omega\tau, \quad s_\tau = \sin \omega\tau, \quad (7)$$

while the sums

$$\begin{aligned} XC &= \sum_j X_j \cos \omega t_j, \\ XS &= \sum_j X_j \sin \omega t_j, \\ CC &= \sum_j \cos^2 \omega t_j, \\ SS &= \sum_j \sin^2 \omega t_j, \\ CS &= \sum_j \cos \omega t_j \sin \omega t_j, \end{aligned} \quad (8)$$

can be evaluated in a single run through the time series, giving a total of $N_f(2N_t + 3)$ trig evaluations for the full periodogram — a factor ~ 2 improvement.

¹ The issue of ‘independent’ frequencies is briefly discussed in Section 6.2.

```

1  __global__ void
2  culsp_kernel(float *d_t, float *d_X, float *d_P,
3              float df, int N_t)
4  {
5
6      __shared__ float s_t[BLOCK_SIZE];
7      __shared__ float s_X[BLOCK_SIZE];
8
9      // Calculate the frequency
10
11     float f = (blockIdx.x*BLOCK_SIZE+threadIdx.x+1)*df;
12
13     // Calculate the various sums
14
15     float XC = 0.f;
16     float XS = 0.f;
17     float CC = 0.f;
18     float CS = 0.f;
19
20     float XC_chunk = 0.f;
21     float XS_chunk = 0.f;
22     float CC_chunk = 0.f;
23     float CS_chunk = 0.f;
24
25     int j;
26
27     for(j = 0; j < N_t; j += BLOCK_SIZE) {
28
29         // Load the chunk into shared memory
30
31         __syncthreads();
32
33         s_t[threadIdx.x] = d_t[j+threadIdx.x];
34         s_X[threadIdx.x] = d_X[j+threadIdx.x];
35
36         __syncthreads();
37
38         // Update the sums
39
40         #pragma unroll
41         for(int k = 0; k < BLOCK_SIZE; k++) {
42
43             // Range reduction
44
45             float ft = f*s_t[k];
46             ft -= rintf(ft);
47
48             float c;
49             float s;
50
51             __sincosf(TWOPI*ft, &s, &c);
52
53             XC_chunk += s_X[k]*c;
54             XS_chunk += s_X[k]*s;
55             CC_chunk += c*c;
56             CS_chunk += c*s;
57
58         }
59
60         XC += XC_chunk;
61         XS += XS_chunk;
62         CC += CC_chunk;
63         CS += CS_chunk;
64
65         XC_chunk = 0.f;
66         XS_chunk = 0.f;
67         CC_chunk = 0.f;
68         CS_chunk = 0.f;
69
70     }
71
72     float SS = (float) N_t - CC;
73
74     // Calculate the tau terms
75
76     float ct;
77     float st;
78
79     __sincosf(0.5f*atan2(2.f*CS, CC-SS), &st, &ct);
80
81     // Calculate P
82
83     d_P[blockIdx.x*BLOCK_SIZE+threadIdx.x] =
84         0.5f*((ct*XC + st*XS)*(ct*XC + st*XS)/
85             (ct*ct*CC + 2*ct*st*CS + st*st*SS) +
86             (ct*XS - st*XC)*(ct*XS - st*XC)/
87             (ct*ct*SS - 2*ct*st*CS + st*st*CC));
88
89     // Finish
90
91 }

```

FIG. 1.— Abridged source for the CULSP computation kernel.

4. CULSP: A GPU LOMB-SCARGLE PERIODOGRAM CODE

4.1. Overview

This section introduces CULSP, a Lomb-Scargle periodogram code implemented within NVIDIA’s CUDA framework. Below, I provide a brief technical overview of CUDA. Section 4.3 then reviews the general design of CULSP, and Section 4.4 narrates an abridged version of the kernel source. The full source, which is freely redistributable under the GNU General Public License, is provided in the accompanying on-line materials.

4.2. The CUDA Framework

A CUDA-enabled GPU comprises one or more streaming multiprocessors (SMs), themselves composed of a number² of scalar processors (SPs) that are functionally equivalent to processor cores. Together, the SPs allow an SM to support concurrent execution of blocks of up to 512 threads. Each thread applies the same computational kernel to an element of an input stream. Resources at a thread’s disposal include its own register space; built-in integer indices uniquely identifying the thread; shared memory accessible by all threads in its parent block; and global memory accessible by all threads in all

blocks. Reading or writing shared memory is typically as fast as accessing a register; however, global memory is two orders of magnitude slower.

CUDA programs are written in the C language with extensions that allow computational kernels to be defined and launched, and the differing types of memory be allocated and accessed. A typical program will transfer input data from CPU memory to GPU memory; launch one or more kernels to process these data; and then copy the results back from GPU to CPU. Executables are created using the `nvcc` compiler from the CUDA software development kit (SDK).

A CUDA kernel has access to the standard C mathematical functions. In some cases, two versions are available (‘library’ and ‘intrinsic’), offering different trade-offs between precision and speed (see Appendix C of NVIDIA 2010). For the sine and cosine functions, the library versions are accurate to within 2 units of last place, but are very slow because the range-reduction algorithm — required to bring arguments into the $(-\pi/4, \pi/4)$ interval — spills temporary variables to global memory. The intrinsic versions do not suffer this performance penalty, as they are hardware-implemented in two special function units (SFUs) attached to each SM. However, they become inaccurate as their arguments depart from the

² Eight, for the GPUs considered in the present work.

$(-\pi, \pi)$ interval. As discussed below, this inaccuracy can be remedied through a very simple range-reduction procedure.

4.3. Code Design

The CULSP code is a straightforward CUDA implementation of the L-S periodogram in its refactored form (equations 6–8). A uniform frequency grid is assumed,

$$f_i = i \Delta f \quad (i = 1, \dots, N_f), \quad (9)$$

where the frequency spacing and number of frequencies are determined from

$$\Delta f = \frac{1}{F_{\text{over}}(t_{N_t} - t_1)} \quad (10)$$

and

$$N_f = \frac{F_{\text{high}} F_{\text{over}} N_t}{2}, \quad (11)$$

respectively. The user-specified parameters F_{over} and F_{high} control the oversampling and extent of the periodogram; $F_{\text{over}} = 1$ gives the characteristic sampling established by the length of the time series, while $F_{\text{high}} = 1$ gives a maximum frequency equal to the mean Nyquist frequency $f_{\text{Ny}} = N_t / [2(t_{N_t} - t_1)]$.

The input time series is read from disk and pre-processed to have zero mean and unit variance, before being copied to GPU global memory. Then, the computational kernel is launched for N_f threads arranged into blocks of size N_b ³; each thread handles the periodogram calculation at a single frequency. Once all calculations are complete, the periodogram is copied back to CPU memory, and from there written to disk.

The sums in equation (8) involve the entire time series. To avoid a potential memory-access bottleneck, and to improve accuracy, CULSP partitions these sums into chunks equal in size to the thread block size N_b . The time-series data required to evaluate the sums for a given chunk are copied from (slow) global memory into (fast) shared memory, with each thread in a block transferring a single (t_j, X_j) pair. Then, all threads in the block enjoy fast access to these data when evaluating their respective per-chunk sums.

4.4. Kernel Source

Figure 1 lists abridged source for the CULSP computational kernel. This is based on the full version supplied in the on-line materials, but special-case code (handling situations where N_t is not an integer multiple of N_b) has been removed to facilitate the discussion.

The kernel accepts five arguments (lines 2–3 of the listing). The first three are array pointers giving the global-memory addresses of the time-series (`d_time` and `d_data`) and the output periodogram (`d_P`). The remaining two give the frequency spacing of the periodogram (`df`) and the number of points in the time series (`N_t`). The former is used on line 11 to evaluate the frequency from the thread and block indices; the macro `BLOCK_SIZE` is expanded by the pre-processor to the thread block size N_b .

Lines 27–70 construct the sums of equation (8), following the chunk partitioning approach described above

³ Set to 256 throughout the present work; tests indicate that larger or smaller values give a slightly reduced performance.

TABLE 1
SPECIFICATIONS FOR THE TWO GPUS USED IN THE BENCHMARKING.

GPU	SMs	SPs	Clock (GHz)	Memory (MB)
GeForce 8400 GS	1	8	1.4	512
Tesla C1060	30	240	1.3	4096

(note, however, that the `SS` sum is not calculated explicitly, but reconstructed from `CC` on line 72). Lines 31–36 are responsible for copying the time-series data for a chunk from global memory to shared memory; the `__syncthreads()` instructions force synchronization across the whole thread block, to avoid potential race conditions. The inner loop (lines 41–58) then evaluates the per-chunk sums; the `\#pragma unroll` directive on line 40 instructs the compiler to completely unroll this loop, conferring a significant performance increase.

The sine and cosine terms in the sums are evaluated simultaneously with a call to CUDA’s intrinsic `__sincosf()` function (line 51). To maintain accuracy, a simple range reduction is applied to the phase `ft` by subtracting the nearest integer [as calculated using `rintf()`; line 46]. This brings the argument of `__sincosf()` into the interval $(-\pi, \pi)$, where its maximum absolute error is $2^{-21.41}$ for sine and $2^{-21.19}$ for cosine (see Table C-3 of NVIDIA 2010).

5. BENCHMARKING CALCULATIONS

5.1. Test Configurations

This section compares the accuracy and performance of CULSP against an equivalent CPU-based code. The test platform is a Dell Precision 490 workstation, containing two Intel 2.33 GHz Xeon E5345 quad-core processors and 8 GB of RAM. The workstation also hosts a pair of NVIDIA GPUs: a Tesla C1060 populating the single PCI Express (PCIe) $\times 16$ slot, and a GeForce 8400 GS in the single legacy PCI slot. These devices are broadly representative of the opposite ends of the GPU market. The 8400 GS is an entry-level product based on the older G80 hardware architecture (the first to support CUDA), and contains only a single SM. The C1060 is built on the newer GT200 architecture (released 2008/2009), and with 30 SMs represents one of the most powerful GPUs in NVIDIA’s portfolio. The technical specifications of each GPU are summarized in Table 1.

The CPU code used for comparison is LSP, a straightforward port of CULSP to ISO C99 with a few modifications for performance and language compliance. The sine and cosine terms are calculated via separate calls to the `sinf()` and `cosf()` functions, since there is no `sincosf()` function in standard C99. The argument reduction step uses an integer cast instead of `rintf()`; this allows the compiler to vectorize the inner loops, greatly improving performance while having a negligible impact on results. Finally, the outer loop over frequency is trivially parallelized using an OpenMP directive, so that all available CPU cores can be utilized. Source for LSP is provided in the accompanying on-line materials.

The Precision 490 workstation runs 64-bit Gentoo Linux. GPU executables are created with the 3.1 release of the CUDA SDK, which relies on GNU `gcc` 4.4 as the host-side compiler. CPU executables are created with Intel’s `icc` 11.1 compiler, using the `-O3` and `-xHost`

optimization flags.

5.2. Accuracy

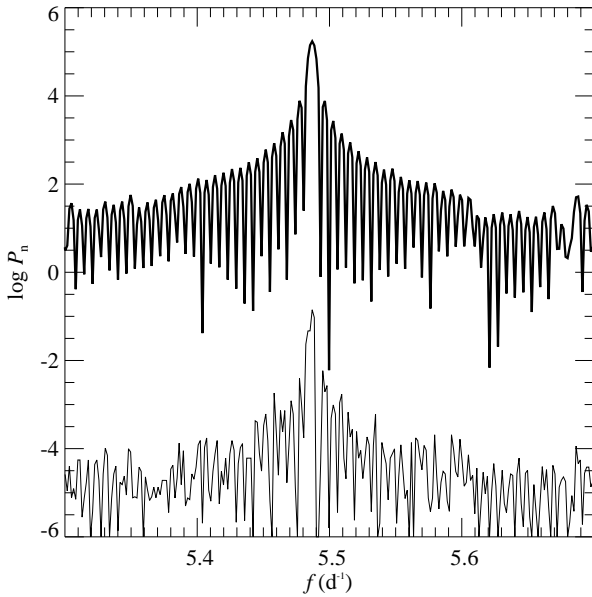


FIG. 2.— Part the L-S periodogram for V1449 Aql, evaluated using the LSP code (thick curve). The thin curve shows the absolute deviation $|P_n^{\text{CULSP}} - P_n^{\text{LSP}}|$ of the corresponding periodogram evaluated using the CULSP code. The strong peak corresponds to the star’s dominant 0.18-d pulsation mode.

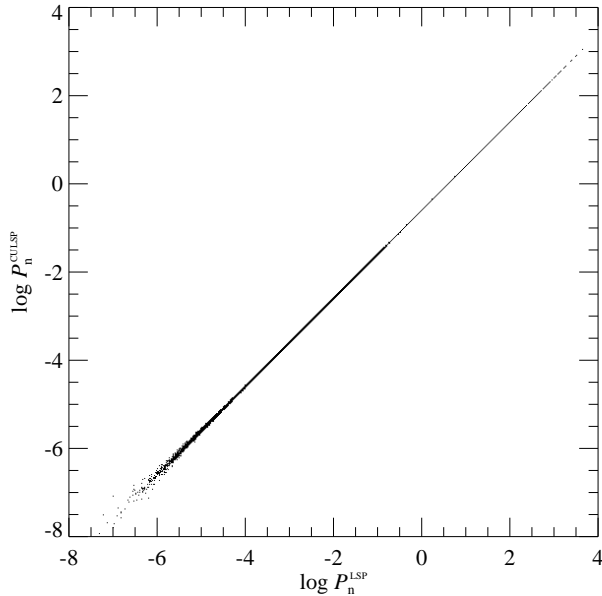


FIG. 3.— A scatter plot of the L-S periodogram for V1449 Aql, evaluated using the LSP (abscissa) and CULSP (ordinate) codes.

As the validation dataset for comparing the accuracy of CULSP and LSP, I use the 150-day photometric time series of the β Cephei pulsator V1449 Aql (HD 180642) obtained by the *CoRoT* mission (Belkacem et al. 2009). The observations comprise 382,003 flux measurements (after removal of points flagged as bad), sampled un-

TABLE 2
PERIODOGRAM CALCULATION TIMES.

Code	Platform	$\langle t_{\text{calc}} \rangle$ (s)	$\sigma(t_{\text{calc}})$ (s)
CULSP	GeForce 8400 GS	570	0.0093
CULSP	Tesla C1060	20.3	0.00024
LSP	CPU (1 thread)	4570	14
LSP	CPU (8 threads)	566	6.9

evenly (in the heliocentric frame) with an average separation of 32 s.

Figure 2 plots the periodogram of V1449 Aql evaluated using LSP, over a frequency interval spanning the star’s dominant 0.18 d pulsation mode (see Waelkens et al. 1998). Also shown in the figure is the absolute deviation $|P_n^{\text{CULSP}} - P_n^{\text{LSP}}|$ of the corresponding periodogram evaluated using CULSP (running on either GPU — the results are identical). The figure confirms that, at least over this particular frequency interval, the two codes are in good agreement with one another; the relative error is on the order of 10^{-6} .

To explore accuracy over the full frequency range, Fig. 3 shows a scatter plot of P_n^{LSP} against P_n^{CULSP} . Very few of the $N_f = 1,528,064$ points in this plot depart to any significant degree from the diagonal line $P_n^{\text{LSP}} = P_n^{\text{CULSP}}$. Those that do are clustered in the $P_n \ll 1$ corner of the plot, and are therefore associated with the noise in the light curve rather than any periodic signal. Moreover, the maximum absolute difference in the periodogram FAPs (equation 3) across all frequencies is 4.1×10^{-5} , which is negligible.

5.3. Performance

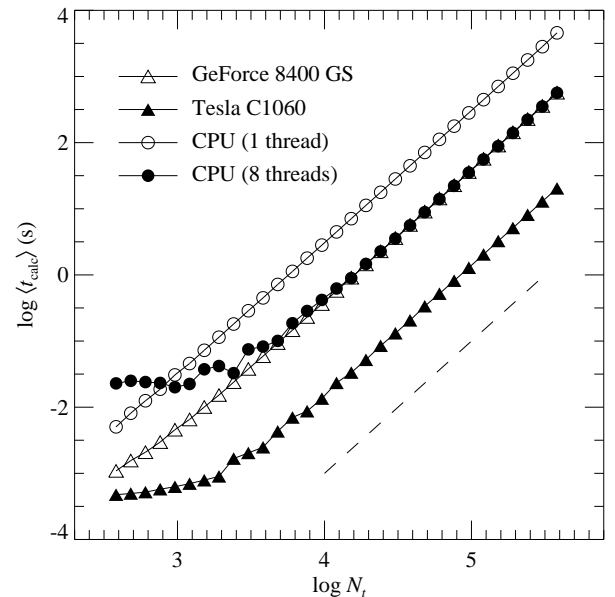


FIG. 4.— Mean calculation times $\langle t_{\text{calc}} \rangle$ for the L-S periodogram, evaluated using the CULSP (triangles) and LSP (circles) codes. The dashed line, with slope $d \log \langle t_{\text{calc}} \rangle / d \log N_t = 2$, indicates the asymptotic scaling of the periodogram algorithm.

Code performance is measured by averaging the V1449 Aql periodogram calculation time t_{calc} over five executions. These timings exclude the overheads incurred by disk input/output and from rectifying light

curves to zero mean and unit variance. Table 2 lists the mean $\langle t_{\text{calc}} \rangle$ and associated standard deviation $\sigma(t_{\text{calc}})$ for CULSP running on both GPUs, and for LSP running with a single OpenMP thread (equivalent to a purely serial CPU implementation), and with 8 OpenMP threads (one per workstation core).

With just one thread, LSP is significantly outperformed by CULSP on either GPU. Scaling up to 8 threads shortens the calculation time by a factor ~ 8 , indicating near-ideal parallelization; nevertheless, the two CPUs working together only just manage to beat the GeForce 8400 GS, and are still a factor ~ 28 slower than the Tesla C1060. Perhaps surprisingly, the latter ratio is *greater* than suggested by comparing the theoretical peak floating-point performance of the two platforms — 74.6 GFLOPS (billions of floating-point operations per second) for all 8 CPU cores, versus 936 GFLOPS for the C1060. This clearly warrants further investigation.

Profiling with the GNU `gprof` tool indicates that the major bottleneck in LSP, accounting for 80% of $\langle t_{\text{calc}} \rangle$, is the `__svml_sincosf4()` function from Intel’s Short Vector Math Library. This function evaluates four sine/cosine pairs at once by leveraging the SSE2 instructions of modern x86-architecture CPUs. Microbenchmarking reveals that a `__svml_sincosf4()` call costs ~ 45.6 clock cycles, or ~ 11.4 cycles per sine/cosine pair. In contrast, thanks to its two special function units, a GPU SM can evaluate a sine/cosine pair in a single cycle (see Appendix G.3.1 of NVIDIA 2010). Scaling these values by the appropriate clock frequencies and processor counts, the sine/cosine throughput on all 8 CPU cores is 1.6 billion operations per second (GOPS), whereas on the 30 SMs of the C1060 it is 39 GOPS, around 23 times faster. Of course, it should be recalled that the GPU `__sincosf()` function operates at a somewhat-reduced precision (see Section 4.4). In principle, the CPU throughput could be improved by developing a similar reduced-precision function to replace `__svml_sincosf4()`. However, it seems unlikely that a *software* routine could ever approach the throughput of the dedicated hardware in the SFUs.

The disparity between sine/cosine throughput accounts for most of the factor ~ 28 performance difference between CULSP and LSP, noted above. The remainder comes from the ability of an SM to execute instructions simultaneously on its SFUs and scalar processors. That is, the sine/cosine evaluations can be undertaken *in parallel* with the other arithmetic operations involved in the periodogram calculation.

Looking now at the memory performance of CULSP, NVIDIA’s `cudaprof` profiling tool indicates that almost all reads from global memory are coalesced, and that no bank conflicts arise when reading from shared memory. Thus, the GPU memory accesses patterns can be considered close to optimal. The combined time spent copying data from CPU to GPU and vice versa is ~ 6 ms on the C1060, and ~ 29 ms on the 8400 GS; while these values clearly reflect the bandwidth difference between the PCIe and PCI slots hosting the GPUs, neither makes any appreciable contribution to the execution times listed in Table 2.

To round off the present discussion, I explore how CULSP and LSP perform with different-sized datasets. The analysis in Section 3 indicates a periodogram workload scaling as $\mathcal{O}(N_f N_t)$. With the number of frequen-

cies following $N_f \propto N_t$ (equation 11), t_{calc} should therefore scale proportionally with N_t^2 — as in fact already claimed in Introduction. To test this expectation, Fig. 4 shows a log-log plot of $\langle t_{\text{calc}} \rangle$ as a function of N_t , for the same configurations as in Table 2. The light curve for a given N_t is generated from the full V1449 Aql light curve by uniform down-sampling.

In the limit of large N_t , all curves asymptote toward a slope $d \log \langle t_{\text{calc}} \rangle / d \log N_t = 2$, confirming the hypothesized N_t^2 scaling. At smaller N_t , departures from this scaling arise from computational overheads that are not directly associated with the actual periodogram calculation. These are most clearly seen in the LSP curve for 8 threads, which approaches a constant $\log \langle t_{\text{calc}} \rangle \approx -1.5$ independent of N_t — perhaps due to memory cache contention between the different threads.

6. DISCUSSION

6.1. Cost/Benefit Analysis

To establish a practical context for the results of the preceding sections, I briefly examine the price vs. performance of the CPU and GPU platforms. At the time of writing, the manufacturer’s bulk (1,000-unit) pricing for a pair of Xeon E5345 CPUs is $2 \times \$455$, while a Tesla C1060 retails for around \$1,300 and a GeForce 8400 GS for around \$50. First considering the C1060, the 50% greater cost of this device (relative to the CPUs) brings almost a factor thirty reduction in periodogram calculation time — an impressive degree of leveraging. However, its hefty price tag together with demanding infrastructure requirements (dedicated PCIe power connectors, supplying up to 200 W), means that it may not be the ideal GPU choice in all situations.

The 8400 GS offers a similar return-on-investment at a much-more affordable price: almost the same performance as the two CPUs at one-twentieth of the cost. This heralds the possibility of budget GPU computing, where a low-end desktop computer is paired with an entry-level GPU, to give performance exceeding high-end, multi-core workstations for a price tag of just a few hundred dollars. Indeed, many desktop computers today, or even laptops, are already well equipped to serve in this capacity.

6.2. Applications

An immediate application of CULSP is analysis of the photometric time series obtained by ongoing satellite missions such as *MOST* (Walker et al. 2003), *CoRoT* (Auvergne 2009), and *Kepler* (Koch 2010). These datasets are typically very large ($N_t \gtrsim 10^5$), leading to a significant per-star cost for calculating a periodogram. When this cost is multiplied by the number of targets being monitored (in the cast of *Kepler*, again $\gtrsim 10^5$), the overall computational burden becomes very steep. Looking into the near future, similar issues will be faced with ground-based time-domain facilities such as Pan-STARRS (Kaiser 2002) and the Large Synoptic Survey Telescope (LSST Science Collaborations and LSST Project 2009). It is hoped that CULSP, or an extension to other related periodograms (see below), will help resolve these challenges.

An additional application of CULSP is in the interpretation of periodograms. Equation (3) presumes that the P_n

at each frequency in the periodogram is independent of the others. This is not necessarily the case, and the exponent in the equation should formally be replaced by some number $N_{f,\text{ind}}$ representing the number of independent frequencies. Horne & Baliunas (1986) pioneered the use of simulations to estimate $N_{f,\text{ind}}$ empirically, and similar Monte-Carlo techniques have since been applied to explore the statistical properties of the L-S periodogram in detail (see Frescura et al. 2008, and references therein). The bottleneck in these simulations are the many periodogram evaluations, making them strong candidates for GPU acceleration.

6.3. Future Work

Recognizing the drawbacks of being wedded to one particular hardware/software vendor, a strategically important future project will be to port CULSP to Open CL (Open Computing Language) — a recently developed standard for programming devices such as multi-core CPUs and GPUs in a platform-neutral manner (see, e.g., Stone et al. 2010). There is also considerable scope for applying the lessons learned herein to other spectral analysis techniques. Shrager (2001) and Zechmeister & Kürster (2009) generalize the L-S periodogram to allow for the fact that the time-series mean is typically not known *a priori*, but instead estimated from the data themselves. The expressions derived by these authors

involve sums having very similar forms to equation (8); thus, it should prove trivial to develop GPU implementations of the generalized periodograms. The multi-harmonic periodogram of Schwarzenberg-Czerny (1996) and the SigSpec method of Reegen (2007) also appear promising candidates for implementation on GPUs, although algorithmically they are rather more-complex.

Looking at the bigger picture, while the astronomical theory and modeling communities have been quick to recognize the usefulness of GPUs (see Section 1), progress has been more gradual in the observational community; radio correlation is the only significant application to date (Wayth et al. 2009). It is my hope that the present paper will help illustrate the powerful data-analysis capabilities of GPUs, and demonstrate strategies for using these devices effectively.

I thank Dr. Gordon Freeman for the initial inspiration to explore this line of research, and the anonymous referee for many helpful suggestions that improved the manuscript. I moreover acknowledge support from NSF *Advanced Technology and Instrumentation* grant AST-0904607. The Tesla C1060 GPU used in this study was donated by NVIDIA through their Professor Partnership Program, and I have made extensive use of NASA's Astrophysics Data System bibliographic services.

REFERENCES

- Aubert D., Teyssier R., 2010, ApJ
 Auvergne M. e. a., 2009, A&A, 506, 411
 Barsdell B. R., Barnes D. G., Fluke C. J., 2010, MNRAS, p. 1200
 Belkacem K., Samadi R., Goupil M., Lefèvre L., Baudin F., Deheuvels S., Dupret M., Appourchaux T., Scuflaire R., Auvergne M., Catala C., Michel E., Miglio A., Montalbán J., Thoul A., Talon S., Baglin A., Noels A., 2009, Science, 324, 1540
 Belleman R. G., Bédorf J., Portegies Zwart S. F., 2008, New Astronomy, 13, 103
 Buck I., Foley T., Horn D., Sugerman J., Fatahalian K., Houston M., Hanrahan P., 2004, ACM Transactions on Graphics, 23, 777
 Cooley J. W., Tukey J. W., 1965, Mathematics of Computation, 19, 297
 Frescura F. A. M., Engelbrecht C. A., Frank B. S., 2008, MNRAS, 388, 1693
 Horne J. H., Baliunas S. L., 1986, ApJ, 302, 757
 Kaiser N. e. a., 2002, in Tyson J. A., Wolff S., eds, SPIE Conf. Ser. 4836 p. 154
 Koch D. G. e. a., 2010, ApJ, 713, L79
 Lomb N. R., 1976, Ap&SS, 39, 447
 LSST Science Collaborations and LSST Project 2009, The LSST Science Book, Version 2.0
 Lyne A., Hobbs G., Kramer M., Stairs I., Stappers B., 2010, Science, 329, 408
 Nguyen H., 2007, GPU Gems 3. Addison-Wesley Professional
 NVIDIA 2010, NVIDIA CUDA Programming Guide 3.1
 Owens J. D., Luebke D., Govindaraju N., Harris M., Krüger J., Lefohn A. E., Purcell T. J., 2005, in Eurographics 2005: State of the Art Reports, p. 21
 Pharr M., 2005, GPU Gems 2. Addison-Wesley Professional
 Press W. H., Rybicki G. B., 1989, ApJ, 338, 277
 Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, Numerical Recipes in Fortran, 2 edn. Cambridge University Press, Cambridge
- Rani B., Gupta A. C., Joshi U. C., Ganesh S., Wiita P. J., 2010, ApJ, 719, L153
 Reegen P., 2007, A&A, 467, 1353
 Rost R. J., 2006, OpenGL Shading Language, 2 edn. Addison-Wesley Professional
 Scargle J. D., 1982, ApJ, 263, 835
 Schive H., Tsai Y., Chiueh T., 2010, ApJS, 186, 457
 Schwarzenberg-Czerny A., 1996, ApJ, 460, 107
 Schwarzenberg-Czerny A., 1998, MNRAS, 301, 831
 Shrager R. I., 2001, Ap&SS, 277, 519
 Simpson E. K., Baliunas S. L., Henry G. W., Watson C. A., 2010, MNRAS, p. 1209
 Stone J. E., Gohara D., Shi G., 2010, Computing in Science and Engineering, 12, 66
 Sturrock P. A., Buncher J. B., Fischbach E., Gruenwald J. T., Javorsek II D., Jenkins J. H., Lee R. H., Mattes J. J., Newport J. R., 2010
 Thompson A. C., Fluke C. J., Barnes D. G., Barsdell B. R., 2010, New Astronomy, 15, 16
 Waelkens C., Aerts C., Kestens E., Grenon M., Eyer L., 1998, A&A, 330, 215
 Walker G., Matthews J., Kuschnig R., Johnson R., Rucinski S., Pazder J., Burley G., Walker A., Skaret K., Zee R., Grocott S., Carroll K., Sinclair P., Sturgeon D., Harron J., 2003, PASP, 115, 1023
 Wayth R. B., Greenhill L. J., Briggs F. H., 2009, PASP, 121, 857
 Zechmeister M., Kürster M., 2009, A&A, 496, 577