

Scalable Photorealistic Rendering of Complex Scenes

Alan Heirich & James Arvo
Department of Computer Science &
Center for Advanced Computing Research
California Institute of Technology, Pasadena CA, 91125

August 30, 1996

Abstract. Photorealistic rendering of complex scenes poses computational demands as great as those of any large scale scientific or engineering calculation. Just as scientific calculations have benefited from access to scalable computing systems so too can photorealistic rendering. This paper describes an application of scalable parallel processors to photorealistic rendering of complex scenes by Monte Carlo path tracing. The application uses scalable implementation methods in order to achieve good performance on large numbers of computers and on models which require large amounts of data. The implementation is a message driven concurrent pipeline which employs a diffusion algorithm for dynamic load balancing. The application can be extended to partition extremely large models across physically distributed memory as well as to perform out-of-core calculations.

1 Introduction

In recent years scalable parallel processors (SPPs) have become readily available to solve computationally intensive problems in science and engineering. Experience with these applications has shown that for certain classes of problems these SPPs routinely achieve speedups close to a factor of P using P computers [5, 15]. Photorealistic rendering of complex scenes poses computational challenges that rival even the largest of these scientific and engineering calculations. It is therefore reasonable to explore how this problem can achieve similar speedups from SPPs.

Amdahl's law states (correctly) that only concurrent computations demonstrate significant scalability on parallel computer systems [2, 12]. Therefore it is not surprising that the first successful applications of SPP technology have been to solve large scale problems in chemistry and physics that involve concurrent phenomena. The natural world abounds with such phenomena. To cite just one example, the process of human vision encompasses a concurrent flow of photons through space, and concurrent computations within the retina of the eye, optic nerve, and visual cortex of the brain. Concurrent processing is a prerequisite to real time vision.

Photorealistic rendering can be expressed as a problem of modeling photon transport through a geometric environment. Ray tracing algorithms directly simulate transport by following the paths of individual photons [22]. A full simulation of this kind can be extremely expensive since the number of paths required to produce a high quality

image can be enormous, and each path can require a significant amount of geometry processing. Path tracing algorithms reduce the number of paths substantially [16]. Despite this improvement a high quality rendering task implemented by path tracing can require consume many hours or even days of computing time on a high performance workstation.

This paper describes a Monte Carlo implementation of path tracing on a parallel computer system. A Monte Carlo method is well suited to parallel implementation because it is inherently concurrent. It is particularly convenient for modeling arbitrary bidirectional reflectance distribution functions (BRDFs) and thus is appropriate for models with directionally diffuse reflective surfaces. In this paper we describe software implementation methods and algorithms that achieve scalable performance on large numbers of computers and accommodate models containing large numbers of geometric primitives. As a result, the implementation can render models that are too large to fit within the memory of any individual computer, and also models that are larger than the space available in all of the computers combined. The approach works on a range of scalable computing systems available at the present time and in the foreseeable future. In particular we have demonstrated the approach on networks of workstations, massively parallel computers, and uniprocessors.

The remainder of this paper is organized as follows. The next section describes the Monte Carlo path tracing algorithm. For completeness we give a brief overview of path tracing with a moderately efficient Monte Carlo sampling strategy and show how it's natural recursion can be expressed more conveniently as an iteration. Following this we describe software implementation methods by which we achieve scalable performance. This performance is demonstrated on a set of images using a diffusion algorithm to perform dynamic load balancing [9, 13, 14]. A set of mapping and load balancing strategies is described for models that are partitioned among computers. Measurements are presented to evaluate the performance of these strategies on extremely large models. Simple extensions to out-of-core computations are discussed, and the paper concludes with a summary and discussion.

2 Monte Carlo path tracing

The radiance (brightness) L_p leaving a point p on a diffusely reflecting surface can be computed from the irradiance (density of incident power) Φ_p at p , which in turn can be estimated from the radiance L_{q_i} arriving from other points q_1, q_2, \dots, q_n in the environment. The radiance arriving from these other points may be estimated by path tracing.

Consider figure 1(a) which shows rays of light traveling indirectly from a luminaire l to a point p in the image plane. The irradiance Φ_A at a selected point A is a sum of the irradiance $\hat{\Phi}_A$ due to direct illumination from light sources and the irradiance due to light that is reflected from surrounding surfaces, such as B, C , and D . Expressing irradiance as a sum of direct and indirect components is an application of the superposition principle, and is extremely useful for Monte Carlo calculations.

For simplicity in the following discussion we shall assume a single light source l . Thus, $\hat{\Phi}_A$ can be expressed as

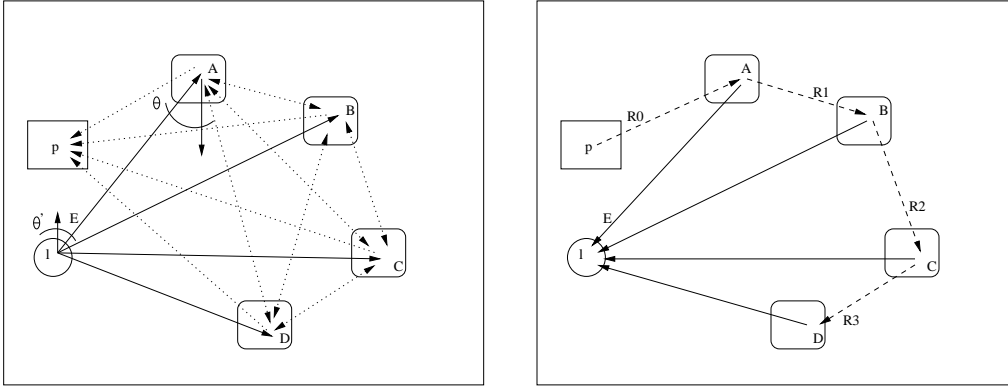


Figure 1: *Computing diffuse interreflection by path tracing. The objective is to compute the radiance L_p at a point p in the image plane. Left figure (a) illustrates light from a luminaire l scattering at points $A, B, C,$ and D on diffusely reflecting surfaces and reaching the image plane at p . Right figure (b) illustrates the path tracing algorithm to estimate the radiance L along a single path. Each step in the path is determined by shooting a ray in a random direction to reach a new point p_i . At each p_i the direct irradiance $\hat{\Phi}_{p_i}$ is computed, attenuated appropriately, and added to L_p . This procedure is typically repeated for many paths.*

$$\hat{\Phi}_A = \int_{\Omega} L_E \cos \theta d\omega, \quad (1)$$

where θ is the angle between a ray E from l to A and the surface normal at point A , and L_E is the radiance (in W/m^2sr) of E . This quantity is integrated over the solid angle Ω subtended by the surface of l , resulting in the irradiance (in W/m^2). The differential solid angle $d\omega$ about a direction E is related to the surface of l by

$$d\omega = \frac{\cos \theta'}{r^2} dA, \quad (2)$$

where θ' is the angle between E and the surface normal of l , dA is the differential surface area of l , and r is the length of E .

To examine the effect of diffuse scattering assume the irradiance Φ_B is known at point B . The associated light energy is scattered by the diffuse surface at B , resulting in equal radiance in all directions. Each of the scattered rays has radiance L_{R_1} equal to $\alpha_B \Phi_B / \pi$, where α_B is a monochromatic reflectance coefficient at B . In this example the point B is visible to the point A in the direction ω . The radiance L_{R_1} , integrated over the solid angle subtended from A by the object containing B , contributes to the irradiance Φ_A at point A . The irradiance $\hat{\Phi}_A$ due to direct illumination from l can be calculated according to (1). The total irradiance Φ_A is the sum of $\hat{\Phi}_A$ and all indirect contributions (such as L_{R_1}),

$$\Phi_A = \hat{\Phi}_A + \int_{\Omega} L_R \cos \theta_A d\omega. \quad (3)$$

Here Ω represents a unit hemisphere above the surface at A . As in (1) the angle θ_A represents the angle of incidence (i.e., the angle between incoming ray R and the surface normal at A). Similarly $d\omega$ represents a differential solid angle, which corresponds to

an area on the surface of the sampling hemisphere. The irradiance at point A , in terms of the irradiance of surrounding Lambertian surfaces, can be expressed

$$\Phi_A = \hat{\Phi}_A + \int_{\Omega} \left(\frac{\alpha_B}{\pi} \Phi_B \right) (\cos \theta_A d\omega) \quad (4)$$

where the location of any point B is a function of A and the direction ω . Specifically, it is a point that is visible to A in the direction ω . (This integral is evaluated over many such points B).

In Monte Carlo path tracing the integral is estimated by tracing individual paths through a geometric environment (see figure 1) and then summing these individual contributions. At each point p along a path one or more sample points q_1, q_2, \dots, q_n from the surrounding environment are used to estimate the irradiance at p , denoted by Φ'_p . The points q_i are random variables; they are the points visible to p in n randomly chosen directions $\omega_1, \omega_2, \dots, \omega_n$ distributed over the hemisphere above p . Thus, we have the estimator

$$\Phi'_p \approx \hat{\Phi}'_p + \frac{1}{\pi n} \sum_{i=1}^n \alpha_{q_i} \Phi'_{q_i} \cos \theta_i \quad (5)$$

It is generally more efficient to sample according to a cosine distribution over the hemisphere, which increases the density of sampling near the pole. This reduces the variance of the estimator and eliminates the factor of $\cos \theta$ from the samples. A cosine distribution is obtained by sampling the unit disk centered at p and projecting upward (along the surface normal at p) to the unit hemisphere. Because $\int_{\Omega} \cos \theta d\omega = \pi$, equation (5) becomes

$$\Phi'_{p_i} \approx \hat{\Phi}'_p + \frac{1}{n} \sum_{i=1}^n \alpha_{q_i} \Phi'_{q_i} \quad (6)$$

The objective of the path tracing procedure is to estimate the radiance L_p at an image point p . Equation (6) describes a recursive estimator, in which each Φ'_{q_i} is written in terms of a sum of several estimated Φ'_{q_i} . A direct implementation of (6) leads to a branching recursion which is well suited to uniprocessor implementation.

Such an implementation leads to practical difficulties on parallel computers. It is difficult in general to ensure that the result of a recursion is returned to its source of origin when data is partitioned among computers, or when work has been moved as a result of dynamic load balancing. It is far more convenient in a parallel context to implement an iterative strategy and this is what we have done. In this iteration rays follow a series of paths that are equivalent to the branching recursion of (6). Rays move forward along the paths until they reach a maximum path length. Along the path the product I_R of a series of α_{p_i} is computed which represents the accumulated attenuation of radiance along the path. At each point p_i the direct irradiance $\hat{\Phi}'_{p_i}$ is attenuated by this product and added to Φ'_p . The final quantity L_p is directly proportional to Φ'_p .

In the iterative path tracing procedure paths start at the eye point and flow into the scene, with successive hops potentially computed on different computers. We shall denote a ray by an ordered-triple (p, ω, I) , where p is the origin, ω is the direction, and I is the importance. Importance is a dimensionless scalar (per color channel) that indicates the fraction of emitted radiance that will contribute to a given pixel. Pixels of

the image are incrementally refined by the loop shown below, in which rays are taken from the queue, processed, and additional rays that were spawned are added back to the queue.

Rays are initially entered into the queue as (p_0, ω_k, I_k) , where p_0 is the eye point, $\omega_1, \omega_2, \dots, \omega_s$ are s directions from p_0 through a given pixel, and I_1, I_2, \dots, I_s are the weights according to some filter kernel; for example, $I_k = 1/s$ for a simple box filter. Each ray also carries with it the index of an associated pixel, which is either the pixel through which it was shot, or the pixel index of the ray that spawned it. When a light source is hit by a ray, a contribution is made to the associated pixel. The main loop for processing rays is as follows:

1. **Cast a ray:** Take ray (p, ω, I) from the queue and find the first point of intersection q on surface S . If the ray is a *shadow ray* then add $I\varepsilon/\pi$ to the associated pixel, where ε is the emissive power of the surface S .
2. **Estimate direct illumination:** Set $I' \leftarrow I\alpha_q/n$, where α_q is the diffuse reflectivity at q . Issue shadow rays $(q, \omega_i, I'A_i \cos \theta_i/r_i^2)$ for $i = 1, 2, \dots, n$, where $\omega_1, \omega_2, \dots, \omega_n$ are directions pointing from q toward n stratified points over the surface of the light source, and A_i is the surface area of stratum i .
3. **Estimate indirect illumination:** If the maximum path length has not been exceeded, set $I'' \leftarrow I\alpha_q/m$ and issue indirect illumination rays (q, ω_j, I'') for $j = 1, 2, \dots, m$, where $\omega_1, \omega_2, \dots, \omega_m$ are directions stratified over the outgoing hemisphere above the point q .
4. **Estimate specular component:** If the surface has a specular component and the maximum path length has not been exceeded, issue the ray $(q, \omega', I\beta_q)$, where β_q is the specular reflectivity of surface S , and ω' is the direction of mirror reflection.

The values of n and m may vary with the depth of the path, the type of surface, or other factors. If $n = 1$ and $m = 1$, then the algorithm corresponds to the path tracing approach introduced by Kajiyama [16].

3 A message driven computation with diffusive load balancing

Whenever a problem can be solved by a modest number of computers it is usually convenient to employ a symmetric multiprocessor (SMP) or other computer system with shared memory. Programming these systems is relatively straightforward because processes can share data and synchronize easily. Programming is more difficult on computer systems with physically distributed memory because communication and synchronization must be managed explicitly. As a matter of architectural necessity the largest parallel computer systems have these properties as do all scalable commodity network configurations [1, 3, 6, 21].

In order to execute a parallel program efficiently it is necessary to maximize concurrency in the program and to distribute the computation evenly among all of the computers. Concurrency may be limited by explicit aspects of the algorithm being employed, for example by explicit requirements for synchronization. Concurrency can

also be limited by artifacts of the implementation, for example as a result of using synchronous communication.

Concurrent pipelines are a useful model for achieving efficient parallelism in iterative computations. A pipeline is constructed by breaking a monolithic iteration into a series of stages. If sufficient storage is available to hold the result of each stage then available concurrency, and therefore computational throughput, can be increased considerably, limited only by the granularity of the stages and the availability of computers to evaluate them. Individual stages can be parallelized to arbitrary degrees. Pipelines avoid synchronization between stages and as a result eliminate the need for synchronous communication.

Message driven computation is a term that is sometimes used to describe a method for implementing concurrent pipelines on parallel computer systems with distributed memory [3, 4, 10]. When a computer executes one stage of a pipeline it stores along with the result of that stage the identity of the next stage which is to process that result. This allows the next stage to be executed by any computer in the system. The result can be transferred to another computer deterministically or as a consequence of a dynamic load balancing operation. Such a strategy makes efficient use of a parallel computer system because the time spent in communication between pipeline stages can be overlapped with the time spent executing those stages. This strategy is effective for hiding a wide range of communication latencies and therefore can be effective on hardware platforms with widely varying communications characteristics. At one extreme this includes platforms with shared memory in which communication latency is virtually zero. At an opposite extreme this includes low cost clustered systems connected by high latency Ethernet.

We have implemented a Monte Carlo path tracing iteration as a message driven concurrent pipeline. The stages of this pipeline correspond primarily to casting bundles of rays. A typical bundle size is 16 or 32 rays which gives a good balance between the amount of computation required to execute one pipeline stage (typically over 100,000 instructions) and the cost of scheduling that stage and managing the data (which can be as little as a few hundred instructions, not including any time that might be required for communication). The implementation does not have an explicit thread of control. Instead it consists of an event loop and a series of event handlers each of which corresponds to a single pipeline stage. When messages are transferred between computers they are routed to appropriate handlers according to their message type. An event handler uses the data in the message as input to a pipeline stage, and typically produces a new message for the result of that stage, marked with an identification of the next stage to execute. For the sake of efficiency a message which corresponds to a ray bundle is processed whenever possible on the same computer as its predecessor. This may not be possible, either because of dynamic load balancing or because a geometric model has been partitioned among a set of computers.

In a full path tracing implementation the pipeline stages do more than cast rays. Our implementation contains stages to perform global synchronization, control startup and termination, build the geometric model, define viewpoints, communicate messages to the user, handle exceptions, monitor performance, and perform a variety of mundane housekeeping chores. Taken together these other stages represent only a small fraction of the overall time spent in the computation.

We employ a diffusion algorithm to perform dynamic load balancing of ray casting

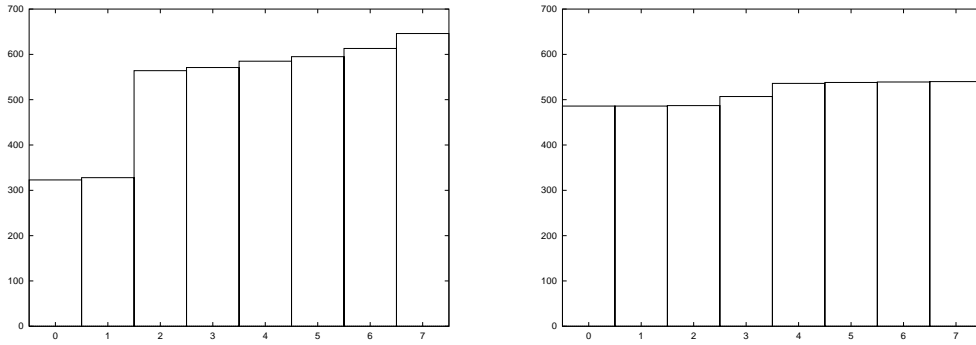


Figure 2: *The result of dynamic load balancing on 8 computers. Left, completion times for a draft quality office image with no load balancing. Completion times varied between 323 and 646 seconds, 50% of the overall elapsed time. Right, completion times for the same image with dynamic load balancing by a diffusion algorithm. These times varied by 54 seconds.*

among computers [9, 13, 14]. Instead of load balancing throughout the computation we have implemented a receiver initiated strategy which becomes active only when a computer has processed all of its messages and has no more pending work. At such a time a computer polls all its neighbors in order to solicit work and requests a specific amount of work as determined by a diffusion calculation [14]. Neighbors respond by supplying the requested amount of work. This strategy produces an overhead from load balancing which is proportional to the number of computers but which does not depend on the length of time spent in computation.

In order to satisfy the requirements of the diffusion calculation it is often necessary to break a large ray bundle into a set of smaller bundles. Rays are chosen as close to termination as possible (i.e., the farthest along the path) since these will spawn the smallest number of subsequent rays and thus conform most closely to the predictions of the diffusion algorithm. When a geometric model is partitioned among computers an additional consideration enters in that portions of the geometric model must be transferred along with the ray bundles.

Despite the variation that can occur in network and computer performance characteristics this implementation should in principle produce scalable performance on most parallel computer systems currently available or available in the near future. The implementation has been tested on a variety of uniprocessors, an Ethernet cluster of Sparcstations, and IBM SP1 and SP2 massively parallel computers. We expect it to run effectively on SMPs where the message passing primitives can be replaced by operations that copy data or pointers to data in memory, and on clusters of SMPs.

4 Scalable performance on a replicated model

The motivation for any use of parallel computers is to achieve as much as P times speedup by using P computers to carry out a calculation. An ultimate goal in the context of computer graphics is to achieve real time rendering. And this is in principle an attainable goal. Nature achieves real time illumination by employing concurrently acting photons and geometric objects. If sufficient computers are available they should in principle be able to simulate this process in real time.

P	Model	Size	T_1	Δ_1	T_2	Δ_2	S	E	T_P	%
1	Office	1	1:05:43	∴	∴	∴	1	1.00	∴	∴
8	Office	1	10:46	5:23	9:00	0:54	7.3	0.91	0:47	0.087
16	Office	1	5:37	3:14	5:01	1:46	13.1	0.82	0:55	0.183
32	Office	1	2:54	1:47	2:45	1:08	23.9	0.75	0:42	0.253
64	Office	1	1:29	0:58	1:17	n/a	51.2	0.80	0:15	0.200
1	Office	9	9:51:27	∴	∴	∴	1	1.00	∴	∴
16	Office	9	50:33	29:12	42:22	8:50	14.0	0.88	5:24	0.127
32	Office	9	25:53	15:40	23:08	n/a	25.6	0.80	4:39	0.201
1	Bath	1	1:20:34	∴	∴	∴	1	1.00	∴	∴
8	Bath	1	15:27	9:43	12:06	1:07	6.7	0.83	2:02	0.168
16	Bath	1	7:51	5:14	5:42	1:35	14.1	0.88	0:40	0.117
32	Bath	1	3:56	2:41	3:51	2:24	20.9	0.65	1:20	0.346

Table 1: *Scalable performance of the implementation with load balancing and replicated data. The office and bathroom models were rendered in two sizes at draft resolution on varying numbers of computers. T_1 is the elapsed time without load balancing, and Δ_1 is the largest discrepancy between times for individual computers. T_2 and Δ_2 are the same quantities when load balancing is in effect. The remaining columns are derived from this data. Speedup S is $T_{(P=1)}/T_2$ where $T_{(P=1)}$ is the value of T_1 when $P = 1$. Efficiency E is S/P . T_P is the total time lost to parallel overheads and is computed as $T_2 - PT_{(P=1)}$. The % column is T_P/T_2 and gives the percentage of the overall elapsed time which is spent in managing concurrency.*

A practical goal for the foreseeable future is to achieve efficient rendering at reasonable rates using readily available hardware. Although parallel computer systems have been constructed from thousands of computers, most readily available systems have a hundred or fewer computers. Whether using a hundred, a thousand, or a million computers, it is necessary to manage concurrency efficiently so that increases in parallel overheads do not result in increases in overall elapsed time. It is likely that real time rendering will require efficient management of fine grained concurrency, in which small amounts of computation are distributed among a large number of computers.

Our design decisions have been driven by these considerations. To meet the goal of efficient concurrency we have employed a concurrent Monte Carlo iteration implemented as a message driven concurrent pipeline. The scaling characteristics of this pipeline are independent of the number of computers on which it executes if communication latencies are properly hidden. Even with imperfect latency hiding the cost of communication can be made acceptably small by adjusting the granularity of the pipeline stages.

In order to approach the ideal of P times speedup using P computers it is essential to have a scalable load balancing mechanism. For this reason we have employed a diffusion algorithm for dynamic load balancing [9]. We have employed a receiver-initiated strategy in order to reduce the number of load balancing events which occur. Diffusive load balancing algorithms have the desirable property of converging in a fixed elapsed time that is independent of P [13]. In a later section of this paper we will propose applying a related technique [14] to the problem of partitioning the geometric model on the computer system in order to achieve similar scalable convergence in this phase of the computation.

In order to evaluate the effectiveness of these strategies we have measured the time required to render the office model at various resolutions and on various numbers of computers. The office model contains 3,256 geometric objects and 3 light sources. For comparison we have also measured the performance on a more complex model of a bathroom. This model contains 18,012 objects and 4 light sources, and has mutually opposing specular surfaces. These measurements are presented in table 1.

The results of table 1 suggest that the present implementation can achieve efficiencies in the neighborhood of 90% when there is sufficient workload to keep all of the computers busy for a substantial length of time. Efficiency drops off somewhat as the amount of computation per processor decreases, until in some cases the computers are spending as much time in maintaining parallelism as they are in performing essential work. This phenomenon occurs commonly in parallel applications. It suggests that the implementation will run efficiently for high quality production images such as the office and bathroom figures. These images required approximately 128 times as many samples as the experiments listed in table 1 for *Size* equal to 1. On 64 computers each of these images might require as much as 2 hours of computation.

We hope to improve these results both for large computations, where we believe efficiency can be boosted consistently above 95%, and for small computations, where we hope to render at near real time rates of less than one second per image. We have identified two principle performance bottlenecks whose removal should help us reach these goals. Neither of these bottlenecks is fundamental but instead represents an area of the implementation which has not yet been tuned for performance.

The first bottleneck is a relatively inefficient termination mechanism. The current termination mechanism suffers from two flaws. Sometimes it allows a computer to

terminate early, and thus become unable to participate in subsequent load balancing events. In addition the mechanism itself scales poorly and nondeterministically. In one instance on 32 computers it was observed to run for more than 30 seconds. Needless to say this greatly increases the overhead spent in managing concurrency for small computations.

The second bottleneck is the use of synchronous and somewhat inefficient communication mechanisms. These mechanisms originated in an implementation which was developed on an Ethernet cluster of workstations. The version of Unix running on that cluster did not support threads at the kernel level and this made it necessary for each computer to periodically poll for incoming message traffic. Since this poll usually fails it results in a significant waste of resources which could otherwise be expended on computation. We have monitored performance on the SP2 which shows this activity taking from 2% to as much as 10% of the available processor time. One reason this is so costly is that the polling requires a kernel call to the `select()` function which entails an expensive switch to kernel mode. In addition to the time wasted in unnecessary polling time is spent in generating message traffic since this requires the same expensive switch to kernel mode.

There are remedies for these problems. Kernel level threads are supported in the newest version of AIX for the SP2 as well as in most of the newest commercial Unix releases. Kernel level threads make it possible to replace polling with an interrupt mechanism which activates a message handling thread whenever traffic arrives from the network. Although time still has to be spent in processing the message our experience has shown that this is negligible in comparison to the time that is currently wasted doing fruitless polling. New software packages are available specifically for the SP2 to allow user mode access to the high performance communication network in order to eliminate the expensive calls to kernel mode services, and explicit support for *active messages* [8, 11]. By a combination of these techniques we expect to eliminate most of the overhead that is currently spent in handling communication.

We are optimistic that by removing inefficiencies in the termination mechanism, and minimizing these communication overheads, we can increase the efficiency on problems of any size above 95%. This would both allow us to improve the performance on large computations, which has already shown acceptable scaling, and to achieve efficient scaling on near real time problems.

5 Rendering very large models

Ray tracing is often called an “embarrassingly parallel” problem because concurrency is abundantly available. When a model is small enough to be replicated on all of the computers then relatively little communication is required during the rendering phase. For example, most of the communication in the experiments reported in table 1 occurred either for dynamic load balancing or to pass the results of sampling to a rasterization process.

It is sometimes impossible to replicate a large model on every computer. In these cases models must be partitioned among the computers. The problem of partitioning is itself challenging (and NP complete) and has consequences for the other phases of the computation. In this section we describe our plans for and initial experiments in addressing these issues. We identify the problems to be solved in mapping and parti-

tioning models among computers. We propose strategies to perform this partitioning, one of which employs a diffusion algorithm which is related to the algorithm we used for load balancing and has similar scaling properties [14]. We describe the impact that this partitioning has upon the other parts of the rendering calculation. An interesting (and unexpected) discovery is that when the necessary modifications have been made to support rendering partitioned models it becomes possible to render models out-of-core, in effect providing user-level control of an extended virtual memory. This approach can be exploited on parallel computers or uniprocessors with equal benefits.

5.1 Mapping and partitioning

The problem of mapping distributed structures onto parallel computers has been studied for many years. This problem was originally posed in terms of mapping a graph of communicating processes onto a network of computers. More recent papers have addressed the analogous problems of mapping data structures [7, 14, 17, 18, 19, 23].

Simple and efficient methods exist to compute a static solution to this problem on a network with regular structure. These methods work by recursively subdividing a data structure into locally contiguous pieces, and assigning the pieces to contiguous regions of the network [17, 23]. For the purposes of rendering these subdivisions can be obtained cheaply by sorting geometric objects according to their relative position in space.

When a network has an arbitrary structure, such as might occur with a network of workstations, or when a problem changes dynamically, such as occurs in virtual reality environments and in animation, these static subdivision methods are inadequate. In recent years a number of papers have begun to explore concurrent *diffusion algorithms* for problems of load balancing and mapping on parallel computers.

Diffusion algorithms can address a number of quadratic minimization problems, including mapping data structures onto parallel computers. The mapping problem is derived from the (NP complete) *graph embedding* problem [20]. In the graph embedding problem the vertices of a relatively large graph, termed the *guest*, are mapped into the vertices of a smaller graph, termed the *host*. The objective of the problem is to find a map which places related guest vertices next to each other in the host, and which maps a similar number of guest vertices to each host vertex.

In the problem of mapping a geometric model onto a set of computers the model corresponds to the guest and the computer network corresponds to the host. A diffusive mapping algorithm can be understood informally as a method to diffuse the geometric elements in a space occupied by a set of computers while satisfying the objective of the graph embedding problem. Specifically, the diffusion algorithm constructs a Euclidean space and assigns regions of the space to computers in the network. It assigns a position in that space to each geometric object. The algorithm works by a simple iteration which moves objects incrementally according to a diffusion process. The algorithm converges to a condition in which related objects have been moved near each other and the distribution of objects is balanced among the computers.

Diffusion algorithms have a number of properties which make them attractive in distributed computing environments. The correctness of these algorithms, their utility and efficiency in parallel computing, as well as a complete bibliography of related work, is given in [14]. The algorithms converge in a fixed amount of elapsed time which is

independent of the number of computers on which they execute. This makes them very attractive for massively parallel implementation. They require no synchronization and communicate only among directly connected computers. They tolerate faults and communication delays without failing. For the purposes of dynamic animation or virtual reality they can reconverge a modified problem rapidly without recomputing it *ab initio*.

A diffusive mapping algorithm can adapt its solution to recognize the availability of different amounts of available memory on different computers. It can also recognize that different geometric objects consume different amounts of storage, and that it is more important for some pairs of objects to be near each other than for others. In order to implement a diffusion algorithm for mapping it is necessary to first construct a graph of nearest neighbor relations among the objects of a geometric model. While this may add substantially to the amount of storage required for the model it only needs to be computed once and thus does not impose a computational burden upon the rendering process.

5.2 *Rendering a partitioned model*

Rendering a partitioned model introduces new considerations for ray casting and load balancing. When a model was replicated, if a ray were cast that did not hit an object, it was safe to assume the ray exited the geometric environment. This is no longer the case when the model is partitioned. A ray that fails to intersect an object must be transferred to a computer containing an appropriate adjacent segment of geometry where it can be cast again.

This may be a straightforward task if each computer is assigned to manage the geometry for a single region of space. Unfortunately in an implementation with load balancing this is not likely to be the case. When uneven workloads develop in a partitioned model it becomes necessary to reassign ownership of partitions. Unless partitions are replicated it is impossible to exchange work in order to perform dynamic load balancing.

A strategy must be designed to allow computers to exchange partitions as well as rays, and to manage rays which are cast between partitions. We have designed such a strategy and implemented its fundamental components. The strategy involves transferring geometric data associated with a partition, i.e., the geometric objects and associated bounding hierarchy, between computers and between computer memory and disk. The strategy works as follows. If computer *A* develops an empty work queue and subsequently requests work from computer *B*, *B* first offers to supply *A* with a copy of its partition. *A* compares the identity of this partition to the identity of its current partition. If these are different then *A* caches its partition data on disk and accepts a copy of *B*'s partition which is transferred as a (very large) message. After that *A* accepts work from *B* in the same way it would if the model were replicated among all of the computers.

During the time that *A* has a partition *k* cached on disk *A* is still the destination for rays which exit other partitions bound for partition *k*. *A* must enqueue these rays and ensure that the entire queue for partition *k* eventually gets evaluated.

We have implemented and tested the fundamental operations of exchanging partitions between computers and caching them to disk. The time required to cache an old partition and acquire a new one is on the order of a few seconds for partitions that are roughly 30 megabytes in size. This suggests that while these transfers should not

occur frequently they are affordable in the context of a large rendering task. It also suggests that out-of-core computations are feasible, in which portions of a geometry are transferred at staged intervals between disk and memory, in effect providing an extremely large virtual memory under the control of the application. We are interested in exploring these issues as well as others as we continue to develop this software.

6 Summary and discussion

This paper describes a design strategy for scalable performance in rendering complex photorealistic scenes on parallel computers. The strategy employs a message driven computation which uses a scalable diffusion algorithm for dynamic load balancing. The implementation does not require synchronization and is relatively insensitive to communication latencies. As a result we expect the implementation to exhibit scalable performance on a range of parallel computer systems, from uniprocessors and symmetric multiprocessors to networks of workstations and tightly coupled massively parallel processors. We have demonstrated it on most of the configurations. On the IBM SP2 we have demonstrated efficiencies in the 80% to 90% range on a variety of problem sizes and numbers of computers, and we anticipate substantial improvements as we continue to refine the code.

The rendering is accomplished by Monte Carlo sampling. Small models are replicated in the memory of every computer. This results in efficient computation and inexpensive dynamic load balancing. We have presented a novel partitioning strategy, based on a diffusion mapping algorithm, which will support models that are too large to replicate on every computer. Initial proof-of-concept simulations and experiments show that under this strategy load balancing will be more expensive because of the need to transfer geometry partitions among computers. This has led, to our surprise, to a strategy for implementing out of core computations in which portions of large models are cycled between disk and memory.

7 Acknowledgments

This project has made extensive use of computer facilities provided by the Caltech Center for Advanced Computing Research, the Cornell Theory Center, and the Argonne High Performance Computing Research Facility. The Argonne HPCRF is funded principally by the U.S. Department of Energy Office of Scientific Computing. Access to the Cornell Theory Center has been provided by the Program of Computer Graphics at Cornell University. Alan Heirich has been supported by Caltech's Center for Advanced Computing Research, the Program of Computer Graphics at Cornell University (NSF Project ASC-9523483, "MRA: Physically and Perceptually-Based Parallel Global Illumination Solutions"), and the Center for Research on Parallel Computation (a National Science Foundation Science and Technology Center, Grant 292-3-51393 under NSF CCR-8809615).

References

- [1] Aboulenen, N., Gjessing, S., Goodman, J. & Woest, P. Hardware support for synchronization in the Scalable Coherent Interface (SCI). *Proc. 8th IEEE Int. Par.*

- Proc. Symp.* (1994).
- [2] Amdahl, G. M. Validity of the single processor approach to achieving large scale computing capabilities. *AFIPS Conf. Proc.* (1967) 483–485.
 - [3] Anderson, T. E., Culler, D. E. & Patterson, D. A. A case for NOW (Networks of Workstations). *IEEE Micro* **15** (1995) 54-64.
 - [4] Athas, W. C. & Seitz, C. L. Multicomputers: message passing concurrent computers. *IEEE Comp.* **21** (1988) 9-24.
 - [5] Baskett, F. & Hennessy, J. L. Microprocessors: from desktops to supercomputers. *Science* **261** (1993) 864–871.
 - [6] Boden, N. J., Cohen, D., Felderman, R. E., Kulawik, A. E., Seitz, C. L., Seizovic, J. N. & Su, W. Myrinet: a gigabit per second local area network. *IEEE Micro* **15** (1995) 29-36.
 - [7] Bokhari, S. Assignment problems in parallel and distributed computing. Boston : Kluwer (1987).
 - [8] Chang, C.-C., Czajkowski, G. & von Eicken, T. Design and performance of active messages on the IBM SP2. Cornell Computer Science Technical Report 96-1572 (1996).
 - [9] Cybenko, G. Dynamic load balancing for distributed memory multiprocessors. *J. Par. Dist. Comp.* **7** (1989) 279-301.
 - [10] von Eicken, T., Culler, D. E., Goldstein, S. C. & Schauer, K. E. Active messages: a mechanism for integrated communication and computation. *Proc. 19th Int. Symp. Comp. Arch.*, Gold Coast, Australia (1992) 256-266.
 - [11] von Eicken, T., Basu, A., Buch, V. & Voegls, W. U-Net: a user level network interface for parallel and distributed computing. *Proc. 15th ACM Symp. Op. Sys. Princ.* (1995) 1-14.
 - [12] Gustafson, J. L. Reevaluating Amdahl's Law. *Comm. ACM* **31** (1987), 532–533.
 - [13] Heirich, A. & Taylor, S. A parabolic load balancing method. *Proc. 24th Intern. Conf. Par. Proc.*, **III**, pp. 192–202. New York : CRC Press (1995).
 - [14] Heirich, A. A scalable diffusion algorithm for dynamic mapping and load balancing on networks of arbitrary topology. *Int. J. Found. Comp. Sci.* (1996), to appear.
 - [15] Hillis, W. D. & Bhogosian, B. M. Parallel scientific computation. *Science* **261** (1993) 856–863.
 - [16] Kajija, J. T. The rendering equation. *Comp. Graph.* **20** (1986).
 - [17] Karypis, G. & Kumar, V. Multilevel graph partitioning schemes. *Proc. 24th Intern. Conf. Par. Proc.*, **III**, pp. 113–122. New York : CRC Press (1995).

- [18] Kung, H. T. & Stevenson, D. A software technique for reducing the routing time on a parallel computer with a fixed interconnection network. In *High speed computer and algorithm organization*, eds. Kuck, Lawrie & Sameh (1977).
- [19] Martin, A. A distributed implementation method for parallel programming. *Inf. Proc.* **80** (1980) 309-314.
- [20] Rosenberg, A. Issues in the study of graph embedding. In *Graph theoretic concepts in computer science*, Lecture Notes in Computer Science **100**, New York : Springer (1981) 150-176.
- [21] Sterling, T. L. The scientific workstation of the future may be a pile-of-pcs. *Comm. ACM* (1996), to appear.
- [22] Whitted, T. An improved illumination model for shaded display. *Comm. ACM* **23** (1980), 343-349.
- [23] Williams, R. D. Performance of dynamic load balancing algorithms for unstructured mesh calculations. *Concurrency: Pract. Exp.* **3** (1991) 457-481.