element substructuring become useless; gathering four elements together will quadruple the total power, so no rays will be saved.

If Monte Carlo methods can achieve $O(N \log N)$ behavior, there is hope that a patch and element system might also. If the number of patches is $O(\log N)$, this will be true for the Hemicube method, and if the number of patches needed for a certain scene is constant (which I suspect may be true), then Wallace's ray tracing method could also be $O(N \log N)$, and the Hemicube method even better. One advantage of the Monte Carlo method is that the optimization seems to happen automatically, so no complicated substructuring or gathering of elements is needed.

6.2.4 Proof of Neccesary Ray Density

Suppose we are generating radiance values for a particular diffuse scene that has been discretized into zones. Assume that we are going to use Monte Carlo ray tracing to estimate the radiance values. In this section I prove that the expected number of rays needed for a solution is $O(N)$, where N is the number of zones, and these zones have certain properties. I will assume a very crude simulation, which is probably much less efficient than the emit and reemit strategy.

Rays will be independently emitted from light sources, each carrying the same amount of power (each of r rays will carry Φ/r power, where Φ is the total power). When a power carrying ray hits a zone, it is probabilistically absorbed or reflected. If reflected, its power is *not* attenuated. Attenuation is implicit because there is a $1 - R$ probability of extinction with each reflection. The ray continues to scatter throughout the environment until it is absorbed. Each ray is absorbed by exactly one zone. The amount of power reflected from the zone can be directly estimated as the power of the rays reflected by that zone. Or we could use the

reflectance multiplied by the total power of absorbed and reflected rays. A third way would be to use the amount of power absorbed as an indirect estimate of the total power reflected. This is possible because the ratio of reflected to absorbed power is simply the ratio of reflectance to absorbance (one minus reflectance). In other words, the ratio of reflected to absorbed power is $R/(1-R)$. I will use this third scheme as our model because it simplifies the mathematics; any one power carrying ray is absorbed exactly once by exactly one surface, but might reflect from many surfaces, or reflect many times from a particular surface.

We would like to show that, given a desired variance bound for our radiance estimates, the expected number of rays traced in a simulation is $O(N)$, where N is the number of zones. If zero area zones are allowed, this will not be true because no rays will ever hit that zone. So we add the restriction that the ratio of the biggest to smallest zone area is bounded. To ensure termination of the physical process, we assume that reflectance is bounded by some number less than one. We also assume the environment we are zoning has some maximum radiance. This value will bound the average radiance of any zone, because the average of a set of values must lie within the range of the set.

Let the following definitions hold:

- N : number of zones.
- A : total area of all surfaces.
- A_i : area of ith zone.
- R_i : reflectance of ith zone.
- R_{max} : maximum reflectance in environment.
- Φ : total emitted power from all zones.

- Φ_i : outgoing reflected power for ith zone.
- L_i : outgoing reflected radiance from ith zone.
- L_{max} : maximum reflected radiance in environment.
- r : number of initial rays emitted by all zones.

We can relate some of these variables immediately. By definition, the radiance of a zone is:

$$L_i = \frac{\Phi_i}{\pi A_i}$$

We assume that the ratio of maximum to minimum zone area is bounded:

$$\frac{A_i}{A_j} < K$$

for all i and j, and for some constant K. This implies that for all i:

$$\frac{A}{KN} < A_i < \frac{KA}{N}$$

Further, we assume that there is some maximum radiance in the scene, L_{max}.

First, let's establish some useful relations. Suppose we have a sum, S, of N identically distributed random variables X_i, where each X_i is a value x with probabilty p and zero otherwise. We can immediately establish:

$$E(S) = E(\sum_{i=1}^{N} X_i) = NE(X_i) = Npx \tag{6.4}$$

and the variance of S is:

$$var(S) = var(\sum_{i=1}^{N} X_i) = Nvar(X_i) = N(E(X_i^2) - E(X_i)^2) = N(px^2 - p^2x^2) \leq Npx^2 \tag{6.5}$$

Initially we send r rays from the emitting zones. We want r large enough so that the variance in our estimate for every L_i, $var(L_i)$, is below some predefined threshold V_0. The

reflected power from the ith zone is:

$$\Phi_i = \frac{R_i}{1 - R_i} \sum_{j=1}^{r} \phi_i^j$$

where ϕ_i^j is the amount of power absorbed from the jth ray by the ith zone. The sum is simply the total power absorbed by the ith zone. Since each ray is absorbed exactly once, the ith zone will either get all of its power (probability p), or none of it (probability $1 - p$). Because all rays are generated independently and according to the same distribution, p depends only on i, and not j, and will thus be denoted p_i. This means the expected value of Φ_i is (from (6.4)):

$$E(\Phi_i) = \frac{R_i}{1 - R_i} p_i \Phi$$

So the expected radiance of the ith zone is:

$$E(L_i) = \frac{1}{\pi A_i} \frac{R_i}{1 - R_i} p_i \Phi$$

Similarly, from (6.5) the variance of the power estimate satisfies the inequality:

$$var(\Phi_i) \leq \frac{R_i^2}{(1 - R_i)^2} p_i \frac{\Phi^2}{r}$$

and similarly:

$$var(L_i) \leq \frac{1}{\pi^2 A_i^2} \frac{R_i^2}{(1 - R_i)^2} p_i \frac{\Phi^2}{r}$$

We can also establish an upper bound for p_i. Since all radinaces are at most L_{max}.

$$E(L_i) = \frac{1}{\pi A_i} \frac{R_i}{1 - R_i} p_i \Phi \leq L_{max}$$

and thus:

$$p_i \leq \pi A_i \frac{1 - R_i}{R_i \Phi} L_{max}$$

So our bound for variance of L_i becomes:

$$var(L_i) \leq \frac{1}{\pi A_i} \frac{R_i}{1 - R_i} L_{max} \frac{\Phi^2}{r}$$

From previous bounds on A_i (all areas at least $A/(KN)$), and R_i we have for all i:

$$var(L_i) \leq \frac{N}{\pi K A} \frac{R_{max}}{1 - R_{max}} L_{max} \frac{\Phi^2}{r}$$

Grouping the constants into a new constant C yields:

$$var(L_i) \leq C \frac{N}{r}$$

To ensure that we estimate all L_i with a variance lower than V_0, we can simply use:

$$r = \frac{CN}{V_0} \tag{6.6}$$

Since each original ray from the light source may reflect off many surfaces, the expected number of *total* rays, the number of intersection calculations needed, as opposed to the number of original power carrying paths, is r times one plus the average number of reflections for each packet. Because reflectivity is bounded, we have:

$$E(r_{total}) \leq r(1 + R_{max} + R_{max}^2 + R_{max}^3 + \cdots) = \frac{r}{1 - R_{max}}$$

So, choosing r according to (6.6) ensures that we will satisfy our target variance condition and:

$$E(r_{total}) \leq \frac{CN}{V_0(1 - R_{max})} = O(N)$$

Thus, the number of rays needed for this Monte Carlo simulation is $O(N)$, where N is the number of zones.

6.2.5 Gauss-Seidel Revisited

One optimization used by Cohen et al. in their progressive refinement method was to keep an ambient term that would make their average error approximately zero (though the absolute error would still be fairly large)[22]. Their steel mill picture was produced by doing a partial

progressive refinement solution with an ambient term, followed by performing the following on each visible surface:

$$\Phi_i = \Phi_i^e + R_i \sum_{j=1}^{N} f_{j \to i} \, \Phi_j$$

This is just one full iteration (for the visible surface) of the Gauss-Seidel method, with a good initial guess. The ambient term is crucial for this to work, or the estimate would be too low. This technique might also work quite well if ray tracing were used, because the number of rays needed for one patch might be much smaller than the number of zones.

6.2.6 Division into Subsystems

Many scenes have subdomains that are disconnected, such as two separate rooms, or subdomains that are partially connected, such as two adjacent rooms with an open door. Some preliminary work on solving these subdomains separately and then linking the solutions has been done by Neumann and Neumann[78] and by Xu et al.[123]. These methods are promising because they will give a speedup for any solution method with time complexity higher than $O(N)$.

6.2.7 Calculation of Direct Lighting in Viewing Phase

Because indirect lighting is soft, a zonal calculation of indirect lighting can use fairly coarse discretization, and a view-dependent solution of the direct lighting can be performed with high accuracy. Both Ward[119] and I[100] have used this idea to produce images. One benefit of doing the direct lighting in the viewing phase is that bump mapping[10] can be used as shown in Figure 6.10. Recently, Chen and Wu have added bump mapping directly to the radiosity phase[18], but their method may behave poorly if sharp shadows fall across a bump mapped surface.

Figure 6.10: Radiosity with image based direct lighting calculation allows proper shading of bump maps.

6.2.8 One Level Path Tracing

Rushmeier developed a method where Kajiya's path tracing is used for primary rays, while secondary (reflected) rays get their radiance values by querying values from a coarse zonal database[92]. Again, this works because the indirect lighting is soft. This technique is appealing because all of the detail of the direct lighting is accounted for, and path tracing is used only while stratification can be maintained. This method brings up the interesting possibility of using an all diffuse zonal phase, while allowing glossy reflection in the viewing phase.

6.3 Zonal Methods for General Reflection Properties

Some of the zonal techniques discussed earlier can be extended to non-diffuse reflection types. The most important application is to scenes that include specular surfaces, but glossy surfaces are sometimes desirable too. In this section these extensions are discussed, and some generalizations of previous methods are developed.

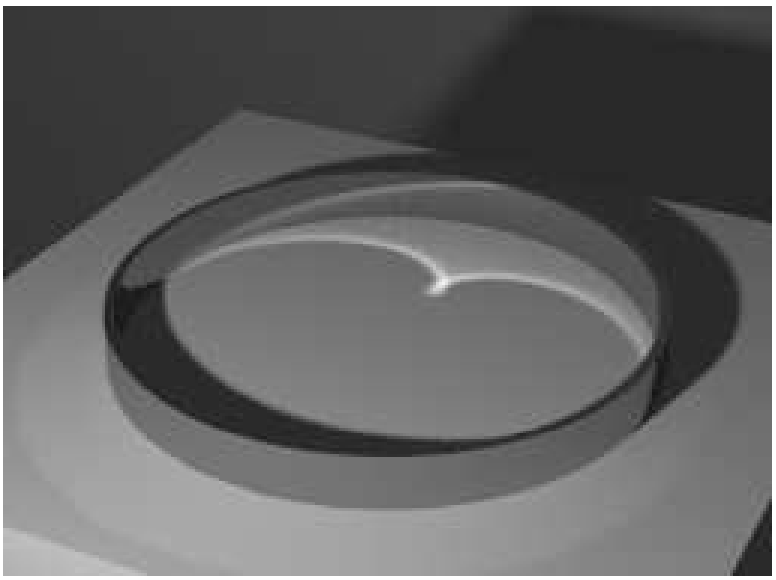

Figure 6.11: Zonal calculations allow for specular transport.

6.3.1 Including Specular Reflection

The simplest method of including specular reflection in a zonal calculation is the *image method*[117, 94]. In the image method, a specular surface is replaced by a hole into a virtual environment. This method works only for planer mirrors, but performs very well for environments that have one important specular surface like a mirror or highly polished floor.

Wallace used distributed ray tracing with a zonal preprocess that allowed highlights and reflections in a radiosity scene, but did not account for most specular transport in the zonal phase[117]. The visual quality of the images produced by this simple method implies that many scenes may not require specular transport in the viewing phase.

Arvo traced rays from the light source in a preprocess that simulated the light transport off specular objects that can cause caustics such as that shown in Figure 6.11[5]. As can be seen in the figure, indirect lighting involving specular surfaces is not always soft. More on this subject can be found in [100]. Heckbert[53] and Watt[120] have also extended Arvo's method.

Malley extended his Monte Carlo power transport method to account for zonal transport by specular surfaces[69]. He did this by allowing power carrying rays to reflect off specular surfaces as shown in Figure 6.12. The colors of specular surfaces can be determined in the viewing phase by standard ray tracing. Sillion and Puech used a similar technique to account for specular reflection, and included subdivision strategies for sampling more heavily where ray paths diverged[103]. An application of the Monte Carlo technique with specular transport is shown in Figures 6.13 and 6.14.

6.3.2 Including General Reflection Functions

The ideas used for specular zonal methods extend nicely to general reflection functions. We can take any set of surfaces, and model them as *reflectors*, rather than absorbers and reemitters as discussed earlier. In the Figure 6.12, the mirror is just acting as a reflector, and the diffuse reflectors are absorbing and then reemitting. There is no reason we cannot make any surface type either a reflector or reemitter, at our discretion.

Any reflectors will not store zonal values and must be assigned colors in the viewing phase. In Figure 6.15, the block on the right is a reflector, and its final colors are determined with path tracing.

Any non-diffuse reflectors can have zonal values, as long as each incoming power packet adds to a power *distribution function* that will be reemitted. In the viewing stage, this distribution can be queried with results depending on viewer position. The distribution functions could be stored in a Hemicube as done by Immel et al.[56], as spherical harmonics as done by Cabral et al.[16], or in hemispherical tables as done by Hall and Rushmeier[42] and Le Saec and Schlick[95].

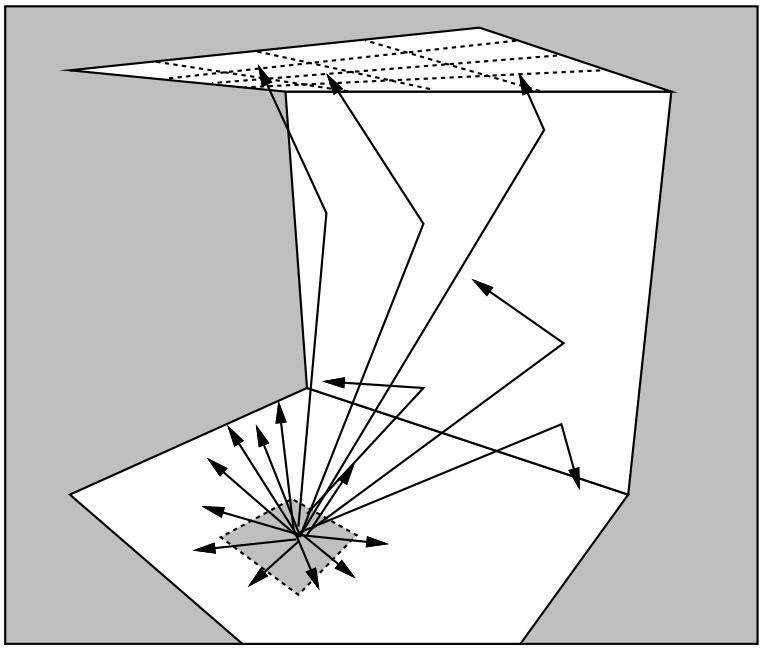

Figure 6.12: Monte Carlo emission of energy with specular reflection.

Figure 6.13: Zonal calculations allow for specular transport.

Figure 6.14: Zonal calculations allow for lighting through glass.

Figure 6.15: Block on right is reflector (not zonal). The rest of the scene is made up of reemitters.

The Monte Carlo method could be used by generating outgoing power rays according to the shape of the unemitted power function.

6.4 Summary

Zonal methods partition the surfaces in a scene into zones which are assumed to be of approximately equal radiance. In the case where all surfaces are diffuse, the radiances of all the zones may be written as a set of simultaneous linear equations and solved by conventional means. It is also possible to bypass forming this system of equations by doing an explicit physical simulation of light transport. This simulation approach can be straightforwardly generalized to scenes that include non-diffuse surfaces. A method was introduced that allows some surfaces to be broken into zones, while others are not. Some time complexity issues were discussed that make ray tracing based energy transport appear potentially better than other zonal strategies.