Rendering and Visualization in Affordable Parallel Environments

Dirk Bartz†
WSI/GRIS
University of Tübingen

Claudio Silva‡, Bengt-Olaf Schneider§
IBM T.J. Watson Research Center

Abstract
The scope of this full-day tutorial is the use of low and medium-cost parallel environments (less than US $ 60K) for high-speed rendering and visualization. In particular, our focus is on the parallel graphics programming of multi-processor PCs or workstations, and networks of both.

The current technology push in the consumer market for graphics hardware, small multiprocessor machines, and fast networks is bound to make all of these components less expensive. In this tutorial, attendees will learn how to leverage these advances in consumer hardware to achieve faster rendering by using parallel rendering algorithms, and off-the-shelf software systems.

This course will briefly touch the necessary tools to make basic use of this technology: parallel programming paradigms (shared memory, message passing) and parallel rendering algorithms (including image-, object-, and time-space parallelism). Advantages and issues of the different methods will be discussed on several examples of polygonal graphics and volume rendering.

Preliminary Tutorial Schedule

Part One: Foundations

Introduction (Bartz/15 minutes)
Personal Workstations (Schneider/45 minutes)
Parallel Architectures (Bartz/30 minutes)
Parallel Programming (Bartz/60 minutes)

Part Two: Rendering

Parallel Polygonal Rendering (Schneider/45 minutes)
Parallel Volume Rendering (Silva/45 minutes)

Part Three: Case Studies

The PVR System (Silva/30 minutes)
Building a Linux-based Parallel Machine (Schneider, Silva/30 minutes)
Q+A (15 minutes)

† Email: bartz@gris.uni-tuebingen.de
‡ Email: csilva@watson.ibm.com
§ Email: bosch@us.ibm.com
PART ONE
Foundations

1. Introduction
This tutorial gives an introduction into the programming of a variety of affordable parallel environments for parallel rendering and scientific visualization. In our case, we define affordable as less than US$ 60,000. PCs are covered as well as workstations; polygonal rendering as well as direct volume rendering.

Overall, the tutorial is organized in three parts. The first part discusses foundations of parallel environments. We discuss Personal Workstations - based on a PC architecture, architectures of multi-processor workstations, and parallel programming using the message-passing and the thread paradigm. The second part introduces into parallel rendering techniques. Specifically, we cover parallel polygonal rendering and parallel direct volume rendering.

In the last part of our tutorial two case studies will be presented. The first case study describes PVR, a parallel rendering system exploiting the message-passing programming paradigm. In the second case study, we discuss a LINUX-based parallel graphics environment. However, the material of the last case study could not be included at press deadline please check the following URL for recent updates of our tutorial:
http://www.gris.uni-tuebingen.de/~bartz/EG_Tutorial/

Our Eurographics’98 tutorial concludes with a question and answer session.

2. Personal Workstations
The advent of powerful processors and robust operating systems for PCs has sparked the creation of a new type of compute platform, the Personal Workstation (PWS). Several vendors, including Compaq, HP, and IBM, sell systems that are targeted at market segments and applications that till only a few years ago were almost exclusively the domain of UNIX-based technical workstations. Such applications include mechanical and electrical CAD, engineering simulation and analysis, financial analysis, and digital content creation (DCC). PWSs are rapidly adopting many features from UNIX workstations, such as high-performance subsystems for graphics, memory, and storage, as well as support for fast and reliable networking. This development creates the opportunity to leverage the lower cost of PWSs to attack problems that were traditionally in the domain of high-end workstations and supercomputers. We will start with an overview of the state of the technology in PWSs and their utility for building parallel rendering systems. Then we will discuss how to improve parallel rendering performance by enhancing PWS subsystems like disks or network connections like disks or

2.1. Architecture
In accordance with the intended application set, PWSs constitute the high-end of the PC system space. Figure 1 shows the architecture of a typical Personal Workstation.

![Architecture of a PWS.](image)

The system contains one or two Pentium II processors, large L2 caches (up to 512 KBytes) and main memory (32 MBytes up to several GBytes). If configured with multiple CPUs, the system acts as a symmetric multiprocessor (SMP) with shared memory. As is well known, shared memory architectures have only limited scalability due to finite access bandwidth to memory. Current PWSs only support dual-processor configurations.

The chipset connects the main processor(s) with other essential subsystems, including memory and peripherals. Among the techniques employed to improve the bandwidth for memory accesses are parallel paths into memory and faster memory technologies, e.g. Synchronous DRAM (SDRAM). Intel has announced that its next generation processor will use Rambus (RDRAM) technology to increase the available memory bandwidth.

The graphics adapter is given a special role among the peripherals due to the high bandwidth demands created by 3D graphics. The Accelerated Graphics Port (AGP) provides a high-bandwidth path from the graphics adapter into main memory. The AGP extends the basic PCI bus protocol with higher clock rate and special transfer modes that are aimed at supporting the storage of textures and possibly z-buffers in main memory, thus reducing the requirements for dedicated graphics memory.

The graphics adapter itself supports at least the OpenGL functionality for triangle setup, rasterization, fragment processing as well as the standard set of 2D functions supported by Windows. Currently, most low-end and midrange graphics adapters rely on the CPU to perform the geometric processing functions, i.e. tessellation of higher-order primitives, vertex transformations, lighting and clipping. However, a new class of high-end PC graphics adapters...
is emerging that implement the geometry pipeline in hardware. Hardware-supported geometry operations are important because rasterizers reach performance levels (several million triangles/sec and several 10 million pixels/sec) that cannot be matched by the system processor(s). Also, geometry accelerators can usually provide acceleration more economically than the CPU, i.e. lower $/MFlops, while freeing the CPU for running applications. However, geometry accelerators will only deliver significant improvements to application performance if the application workload contains a large portion of graphics operations. Many applications (and application-level benchmarks) contain only short bursts of graphics-intensive operations.

Balancing the system architecture requires fast disk, e.g. 10,000 rpm SCSI disk drives, and networking subsystems, e.g. 100 Mbit/sec or 1Gbit/sec Ethernet.

### 2.2. Parallel Configurations

For the purposes of parallel rendering we will be considering two forms of parallelism: tightly coupled processors in a SMP configuration (as shown in Figure 1) and a cluster of workstations connected over networks. While in a single-processor machine CPU performance is often the most important factor in determining rendering performance, parallel configurations add specific constraints to the performance of parallel rendering algorithms. For SMP workstations, the performance is affected by memory and disk bandwidth. For workstation clusters, the disk and network bandwidth are the most important parameters influencing the rendering performance. The next section provides concrete values for these parameters.

### 2.3. Performance

To illustrate the performance that can be expected from a PWS we provide approximate performance data in Table 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integer performance</td>
<td>570 MIPS</td>
</tr>
<tr>
<td>Floating point performance</td>
<td>220 MFLOPS</td>
</tr>
<tr>
<td>Memory bandwidth</td>
<td>140 MBytes/sec</td>
</tr>
<tr>
<td>Disk bandwidth</td>
<td>13 MBytes/sec</td>
</tr>
</tbody>
</table>

Table 1: Approximate peak performance data for a Personal Workstation.

We have also conducted measurements of networking performance using various local area network technologies (Table 2). These measurements consisted of transferring large data packets and used the TCP/IP stack that is part of Windows NT 4.0. Note that the observed bandwidth for Gigabit-Ethernet is far below the expected value. A likely source for this shortfall is inefficiencies in the implementation of the TCP/IP stack and the resulting high CPU loads. It is well known that such inefficiencies can result in severe performance degradations and we expect that a better TCP/IP stack would raise the transfer rate.

### 2.4. PWS Market Trends

So far we have reviewed technical characteristics of PWSs. When selecting a workstation platform technical issues are but one factor.

The developments in the PWS market reflect the PWS’s dual-inheritance from Unix workstations and PCs.

As the NT workstation markets matures the price gap between the best performing systems and the systems with best price-performance appears to be closing. This is a known trend know from the desktop PC market which has turned into a commodity market. The PWS market is assuming characteristics of a commodity market with surprising speed, i.e. most products are very similar and have to compete through pricing, marketing and support offerings.

At the same time, PWSs remain different from desktop PCs – and are similar to Unix workstations – in that application performance (in contrast to servicability and manageability) is the primary design and deployment objective. Most purchasing decisions are heavily based on the results in standard and customer-specific application benchmarks.

A particularly interesting question is whether PWSs offer inherently better price-performance than traditional Unix workstations. Over the period that both workstation types participated in the market (1996-1998), NT workstations as a whole have consistently delivered better price-performance than Unix workstations for standard benchmarks. Only recently (mid 1998) Unix workstation are beginning to reach the same price-performance levels. It is unclear whether this constitutes a reversal of the earlier trend or whether the gap will be restored when Intel delivers its next generation processors. Another explanation for the narrowing of

<table>
<thead>
<tr>
<th>LAN Technology</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token Ring 16 Mbit/sec</td>
<td>14-15 Mbit/sec</td>
</tr>
<tr>
<td>Ethernet 10 Mbit/sec</td>
<td>7-8 Mbit/sec</td>
</tr>
<tr>
<td>Ethernet 100 Mbit/sec</td>
<td>90 Mbit/sec</td>
</tr>
<tr>
<td>Ethernet 1 Gbit/sec</td>
<td>120 Mbit/sec</td>
</tr>
</tbody>
</table>

Table 2: Peak bandwidth between Personal Workstations for different LAN technologies.

this gap is that NT workstations are starting to include high-performance subsystems that are required for balanced systems (see below).

2.5. Building Parallel Renderers from Personal Workstations

Parallel rendering algorithms can be implemented on a variety of platforms. The capabilities of the target platform influence the choice of rendering algorithms. For instance the availability of hardware acceleration for certain rendering operations affects both performance and scalability of the rendering algorithm.

Several approaches to implementing parallel polygon rendering on PWSs with graphics accelerators have been investigated by Schneider\textsuperscript{104}. It should be noted that this analysis does not consider pure software implementations of the rendering pipeline; rasterization was assumed to be performed by a graphics adapter.

This is in contrast to software-only graphics pipelines. Such approaches lead to more scalable rendering systems, even though both absolute performance and price-performance are likely to be worse than the hardware-accelerated implementation. In a paper by Whitman\textsuperscript{125}, parallel software renderers have shown close speedup up to 100 processors in a BBN Butterfly TC2000 even though the absolute performance (up to 100,000 polygons/sec) does not match the performance available from graphics workstations of equal or lower cost. However, software renderers offer more flexibility in the choice of rendering algorithms, e.g. advanced lighting models, and the option to integrate application and render more tightly.

Following the conclusions from Schneider\textsuperscript{104} we will now look at the various subsystems in a PWS that may become a bottleneck for parallel rendering. In part, PWSs have inherited these bottlenecks from their desktop PC ancestors. For example, both memory and disk subsystems are less sophisticated than those of traditional workstations. We will also discuss the merit of possible improvements to various subsystems with respect to parallel rendering performance.

Applications and Geometry Pipeline. As pointed out above, CPU portion of the overall rendering time scales well with the number of processors. Therefore, it is desirable to parallelize rendering solutions with a large computational component. Advance rendering algorithms such as advanced lighting algorithms or ray-tracing will lead to implementations that scale to larger numbers of processors.

Processor. Contrary to initial intuition, the performance of CPU and rasterizer does not significantly influence the overall rendering performance. Therefore, parallel rendering does not benefit from enhancements to the CPU, such as by higher clock frequency, more internal pipelines or special instructions to accelerate certain portions of the geometry pipeline. However as stated earlier, faster CPUs may benefit the application's performance.

Memory Subsystem. Currently, memory bandwidth does not limit rendering performance as much as disk and network performance. We expect that memory subsystems will keep increasing their performance over time and retain their relative performance compared to disks and networks. Therefore, more sophisticated memory subsystems, like \textsuperscript{2}, will not improve parallel rendering performance.

Disk Subsystem. The disk subsystem offers ample opportunity for improvements over the standard IDE or SCSI found in today's PWSs. Faster disk subsystems, e.g. SSA\textsuperscript{1} or RAID 0 (disk striping), can be used to alleviate this problem.

Graphics Subsystem. In workstation clusters the use of graphics accelerators with geometry accelerators can be beneficial. For applications with mostly static scenes, e.g. walkthroughs or assembly inspections, the use of retained data structures like display lists can reduce the bandwidth demands on system memory as geometry and lighting calculations are performed locally on the adapter. In SMP machines or for single-frame rendering faster graphics hardware will not provide large rendering speed-ups.

Network. In clusters, a slow network interconnect can become the dominant bottleneck. Increasing the network bandwidth by an order of magnitude will alleviate that problem. As stated above, current shortcomings of the protocol implementations prevent full realization of the benefits of Gigabit-Ethernet under Windows NT. Alternative technologies, like Myrinet\textsuperscript{36} promise higher sustained bandwidth than Ethernet. However, these technologies are either not available under Windows NT or have not yet been developed into a product. Prototype implementations under Unix (Linux) have demonstrated the advantages of such networks.

2.6. Conclusion

As Personal Workstations are emerging as an alternative to traditional workstations for technical applications they are frequently considered as building blocks for affordable parallel rendering.

Even though PWS are used for parallel rendering in at least one commercial rendering package\textsuperscript{43}, its actual implementation is hampered by the lack of efficient networking technologies and insufficient disk performance. Improving these subsystems is possible but will result in more expensive systems, eliminating some of the perceived cost advantage of PWS over traditional workstation.

3. Parallel Architectures

In this Section, we discuss general aspects of parallel environments. Although our tutorial covers PCs and workstations, we will focus in this Section only on workstation
environments. However, most of the information on software aspects (message passing, process communication, and threads) is applicable to all UNIX environments (e.g. Linux).

The following Sections will discuss the different parallel approaches, architectures, and programming models for parallel environments.

3.1. Parallel Approaches

Three basic approaches are available for parallel environments. The first approach connects different computers via a network into a cluster of workstations (or PCs). On each individual computer processes are started to perform a set of tasks, while communication is organized by exchanging messages via UNIX sockets, message passing (e.g. PVM), or - more recently - via the Internet. We call this type a loose coupled system, sometimes referred as a distributed processing system.

The second approach consists of a single computer, which contains multiple processing elements (PE which actually are processors). These processing elements are communicating via message passing on an internal high-speed interconnect, or via memory. This type is called a tight coupled system. In contrast to the first approach, communication is faster, usually more reliable, and - in the case of a shared memory system - much easier to handle. However, depending of the interconnection system, the number of processing elements is limited.

The third basic approach is a fusion of the first two approaches. We generally can combine tight or loose coupled systems into a hybrid coupled system. However, in most cases we will loose the advantages of a tight coupled system.

3.2. Taxonomy

Flynn developed a taxonomy to classify the parallel aspects of the different (more or less) parallel systems. However, this taxonomy actually only applies to tight coupled systems.

Flynn distinguishes two basic features of a system, the instruction stream (I) - which is code execution - and the data stream (D) - which is the data flow. These features are divided into a single (S) or multiple data stream (M). In a single instruction stream, only one instruction can be individually performed by a set of processors, while a multiple instruction stream can perform different instructions at the same time. If we have a single data stream, only this data can be computed or modified at the same time. With a multiple data stream, more than one data element can be processed.

Overall, we have four different types of parallel processing:

- **SISD** - is the standard workstation/PC type. A single instruction stream of a single processor is performing a task on a single data stream.

- **SIMD** - is the massively-parallel, or array computer type. The same instruction stream is performed on different data. Although a number of problems can easily mapped to this architecture (e.g. matrix operations), some problems are difficult to solve with SIMD systems. Usually, these systems cost hundreds of thousands of US$. One of the reasons these machines are not covered by this tutorial.

- **MISP** - is not a useful system. If multiple instructions are executed on a single data stream, it will end up in a big mess. Consequently, there are no computer systems using the MISP scheme.

- **MIMD** - is the standard type of a parallel computer. Multiple instruction streams perform their task on their individual data stream.

3.3. Memory Models

Many aspects of parallel programming depend on the memory architecture of a system, and many problems arise from a chosen memory architecture. The basic question is if the memory is assigned to the processor level, or if the memory is assigned on system level. This information is important for the distribution of a problem to the system. If all memory - except caches - is accessible from each part of the system - memory is assigned on system level, we are talking of a shared memory system. In case the individual processing elements can only access their own private memory - memory is assigned on processor level, we are talking of a distributed memory system. Shared memory systems are further divided into UMA (Uniform Memory Access) systems (not interchangeable with Uniform Memory Architecture), and into NUMA (Non-Uniform Memory Access) systems.

3.3.1. Distributed Memory Systems

In distributed memory systems, the memory is assigned to each individual processor. At the beginning of the processing, the system is distributing the tasks and the data through the network to processing elements. These processing elements receive the data and their task and start to process the data. At some point, the processors need to communicate with other processors, in order to exchange results, to synchronize for periphery devices, and so forth. Finally, the computed results are sent back to the appropriate receiver and the processing element waits for a new task. Workstation clusters fit into this category, because each computer has its individual memory, which is not accessible from its partner workstations within the cluster. Furthermore, each workstation can distribute data via the network.

Overall, it is important to note that communication in a distributed memory system is expensive. Therefore, it should be reduced to a minimum.
3.3.2. Shared Memory Systems

UMA systems contain all memory in a more or less monolithic block. All processors of the system access this memory via the same interconnect, which can be a crossbar or a bus (Figure 2). In contrast, NUMA systems are combined of two or more UMA levels which are connected via another interconnect (Figure 3). This interconnect can be slower than the interconnect on the UMA level. However, communication from one UMA sub-system to another UMA sub-system travels through more than one interconnection stage and therefore, takes more time than communication within one UMA sub-system.

If UMA systems have a better communication, why should we use NUMA systems? The answer is that the possibilities to extend UMA systems are limited. At some point the complexity of the interconnect will rise into infinity, or the interconnect will not be powerful enough to provide sufficient performance. Therefore, a hierarchy of UMA sub-systems was introduced.

3.4. Programming Models

So far, we have introduced different approaches of parallelization (loose coupled or distributed processing, tight-coupled processing, and hybrid models of loose- or tight-coupled processing) and different memory access architectures. In this Section, we add two different paradigms for the programming of parallel environments.

† We are talking of main memory. Processor registers, caches, or harddiscs are not considered as main memory.

3.4.1. Message-Passing

This programming paradigm connects processing entities to perform a joined task. As a matter of principle, each processing entity is an individual process running on a computer. However, different processes can run on the very same computer, especially, if this computer is a multi-processor system. The underlying interconnection topology is transparent from the users point of view. Therefore, it does not make a difference in programming, if the parallel program which communicates using a message-passing library runs on a cluster of workstations, on a distributed memory system (e.g. the Intel Paragon), or on a shared-memory system (e.g. the HP Convex/SPP).

For the general process of using a message-passing system for concurrent programming it is essential to manually split the problem to be solved into different more or less independent sub-tasks. These sub-tasks and their data are distributed via the interconnect to the individual processes. During processing, intermediary results are sent using the explicit communication scheme of message-passing.

Considering the high costs using the network, communication must be reduced to a minimum. Therefore, the data must be explicitly partitioned. Finally, the terminal results of the processing entities are collected by a parent process which returns the result to the user.

Their are several message-passing libraries around. However, most applications are based on two standards, which are explained in Section 4.2 and Section 4.1; the PVM3 library (Parallel Virtual Machine) and the MPI standard (Message Passing Interface).

3.4.2. Threading

A more recent parallel programming paradigm is the thread model. A thread is a control flow entity in a process. Typically, a sequential process consists of one thread; more than one thread enable a concurrent (parallel) control flow. While the process provides the environment for one or more threads - creating a common address space, a synchronization and execution context - the individual threads only build a private stack and program counters. The different threads of a single process communicate via synchronization mechanisms and via the shared memory.

Sometimes the concept of light-weight processes (LWP) is used as a synonym for threads. However, a LWP actually is a physical scheduling entity of the operating system, in a way the physical incarnation of the logical concept of a thread.

In contrast to message passing, threading is only possible on multi-processor systems. Moreover, multi-processor systems need a shared-memory architecture, in order to provide the same virtual address space.

‡ There are some thread models which run on distributed mem-
Besides easy communication and data exchange using the shared memory, switching between different threads is much cheaper/faster than switching between individual processes. This is due to the shared address space, which is not changed during a thread switch.

Basically, there are three different kinds of implementations for threads. There is a user thread model, a kernel thread model, and a mixed model. The user thread model is usually a very early implementation of a thread package. All thread management is handled by the thread library; the UNIX kernel only knows the process, which might contain more than one thread. This results in the situation that only one thread of a process is executed at any particular time. If you are using threads on a single processor workstation, or your threads are not compute-bound, this is not a problem. However, on a multi-processor system, we do not really get a concurrent execution of multiple threads of one process. On the other hand, this implementation model does not require a modification of the operating system kernel. Furthermore, the management of the threads does not require any kernel overhead. In Pthread terminology, this model is called all-to-one-scheduling.

In contrast to user threads, each kernel thread (on Solaris systems a kernel thread is called a light-weight process, on SGI systems a sproc) is known to the operating system kernel. Consequently, each kernel thread is individually schedulable. This results in a real concurrent execution on a multi-processor, which is especially important for compute-bound threads. However, allocation and management of a kernel thread introduces significant overhead to the kernel, which eventually might lead to a bad scaling behaviour. Pthread terminology denotes this model to be one-to-one-scheduling.

As usual, the best solution is probably a mixed model of user and kernel threads. The threads are first scheduled by the thread library (user thread scheduling). Thereafter, the threads scheduled by the library are scheduled as kernel threads. Threads that are not compute-bound (e.g. performing I/O) are preempted by the scheduling mechanism of the library, while only compute-bound threads are scheduled by the kernel, thus enabling high-performance concurrent execution. In Pthread terminology, this model is called the many-to-one or some-to-one scheduling.

### 3.5. Example Architectures

In Table 3, we present an overview of different systems which cost about or less than US$ 60,000. Please note that all price information is selected from the different web sites of the different companies. Therefore, discounts and sales tax are not included.

More or less, we always tried to configure a standard system with a four GB harddisc and 128 MB of main memory. (This may appear not enough memory, but who buys memory from the system vendor anyway?). Prices are in US$ (considering an exchange rate DEM 1.8/US$ 1).

#### Sun Enterprise 450

Figure 4 gives an overview of the Sun Ultra Enterprise 450 architecture. Up to four processors are connected via a crossbar to the UMA memory system and to the I/O system. The processors are managed via the system controller. On Sun workstations/servers, pthreads are available as mixed model implementation (Solaris 2.5 and above).

#### Hewlett-Packard D-class/J-class architecture

In Figure 5, the basic architecture of D-class and J-class architecture of Hewlett-Packard is shown. Up to two processors are connected via the memory bus to the UMA memory system and the I/O system. Similar to this architecture, the K-class servers can connect up to four processors. On HP-UX 10.30, pthreads are available as a kernel model. Older versions implement a user model.
Silicon Graphics Octane architecture

The processor boards of the SGI Octane architecture contain up to two processors and the UMA memory system. These boards are connected via a crossbar with the Graphics system and the I/O system (Figure 6). Pthreads are available for IRIX 6.3 and above, where pthreads are available as patch set for IRIX 6.2. On all implementations, a mixed model is used.

![Figure 6: Basic SGI Octane architecture](image)

Silicon Graphics Origin200 architecture

In contrast to the SGI Octane, no crossbar is used for the Origin200 architecture. The single tower configuration (up to two processors) connects the processors with the UMA memory system and the I/O system via a hub interconnect. For the four processors configuration, a “Craylink” interconnect links two two processors towers system to a Non-Uniform Memory Access (NUMA) system (Figure 7). In the case of the Origin200, a cache-coherent NUMA scheme is implemented, in order to provide a consistent memory view for all processors. Pthreads are available for IRIX 6.3 and above, where pthreads are available as patch set for IRIX 6.2. On all implementations, a mixed model is used.

![Figure 7: Basic SGI Origin 200 architecture](image)
<table>
<thead>
<tr>
<th>Vendor/Model</th>
<th>CPU(s)</th>
<th>[N]UMA</th>
<th>Interconnect</th>
<th>Memory</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun/Enterprise 450</td>
<td>1 @250 MHz</td>
<td>UMA</td>
<td>crossbar @1.6 GB/s</td>
<td>&lt;4 GB</td>
<td>19K</td>
</tr>
<tr>
<td></td>
<td>2 @250 MHz</td>
<td>UMA</td>
<td>crossbar @1.6 GB/s</td>
<td>&lt;4 GB</td>
<td>27K</td>
</tr>
<tr>
<td></td>
<td>4 @250 MHz</td>
<td>UMA</td>
<td>crossbar @1.6 GB/s</td>
<td>&lt;4 GB</td>
<td>42K</td>
</tr>
<tr>
<td></td>
<td>1 @300 MHz</td>
<td>UMA</td>
<td>crossbar @1.6 GB/s</td>
<td>&lt;4 GB</td>
<td>25K</td>
</tr>
<tr>
<td></td>
<td>2 @300 MHz</td>
<td>UMA</td>
<td>crossbar @1.6 GB/s</td>
<td>&lt;4 GB</td>
<td>39K</td>
</tr>
<tr>
<td></td>
<td>4 @300 MHz</td>
<td>UMA</td>
<td>crossbar @1.6 GB/s</td>
<td>&lt;4 GB</td>
<td>66K</td>
</tr>
<tr>
<td>HP/J Class J2240</td>
<td>2 @236 MHz</td>
<td>UMA</td>
<td>bus @960 MB/s</td>
<td>&lt;1 GB</td>
<td>32K</td>
</tr>
<tr>
<td>HP/D Class D370</td>
<td>1 @160 MHz</td>
<td>UMA</td>
<td>bus @960 MB/s</td>
<td>&lt;1 GB</td>
<td>29K</td>
</tr>
<tr>
<td></td>
<td>2 @160 MHz</td>
<td>UMA</td>
<td>bus @960 MB/s</td>
<td>&lt;1 GB</td>
<td>39K</td>
</tr>
<tr>
<td>HP/D Class D380</td>
<td>1 @180 MHz</td>
<td>UMA</td>
<td>bus @960 MB/s</td>
<td>&lt;1 GB</td>
<td>31K</td>
</tr>
<tr>
<td></td>
<td>2 @180 MHz</td>
<td>UMA</td>
<td>bus @960 MB/s</td>
<td>&lt;1 GB</td>
<td>44K</td>
</tr>
<tr>
<td>SGI/Octane SE</td>
<td>1 @225 MHz</td>
<td>UMA</td>
<td>crossbar @1.6 GB/s</td>
<td>&lt;2 GB</td>
<td>24K</td>
</tr>
<tr>
<td></td>
<td>2 @225 MHz</td>
<td>UMA</td>
<td>crossbar @1.6 GB/s</td>
<td>&lt;2 GB</td>
<td>30K</td>
</tr>
<tr>
<td></td>
<td>1 @250 MHz</td>
<td>UMA</td>
<td>crossbar @1.6 GB/s</td>
<td>&lt;2 GB</td>
<td>30K</td>
</tr>
<tr>
<td></td>
<td>2 @250 MHz</td>
<td>UMA</td>
<td>crossbar @1.6 GB/s</td>
<td>&lt;2 GB</td>
<td>42K</td>
</tr>
<tr>
<td>SGI/Origin 200</td>
<td>1 @180 MHz</td>
<td>UMA</td>
<td>hub @1.28 GB/s</td>
<td>&lt;2 GB</td>
<td>16K</td>
</tr>
<tr>
<td></td>
<td>2 @180 MHz</td>
<td>UMA</td>
<td>hub @1.28 GB/s</td>
<td>&lt;2 GB</td>
<td>22K</td>
</tr>
<tr>
<td></td>
<td>4 @180 MHz</td>
<td>ccNUMA</td>
<td>hub/craylink @1.28 GB/s</td>
<td>&lt;4 GB</td>
<td>42K</td>
</tr>
</tbody>
</table>

**Table 3:** Systems overview.
4. Parallel Programming

A. Message Passing

In this part of the tutorial, we briefly introduce two message-passing libraries. First we discuss the Message-Passing Interface library - MPI\textsuperscript{10, 40}, followed by the Parallel Virtual Machine library - PVM\textsuperscript{42, 10}. A comparison of these libraries can be found in an article by G. Geist et al.\textsuperscript{43}. All these papers can be found on the web, either at netlib, or at the respective homepages of the libraries.

4.1. Message Passing Interface - MPI

MPI 1 (1994) (and later MPI 2 (1997)) is designed as a communication API for multi-processor computers. Usually, the functionality of MPI is implemented using a communication library of the vendor of the machine. Naturally, this vendor library is not portable to other machines. Therefore, MPI adds an abstraction level between the user and this vendor library, in order to guarantee the portability of the program code of the user.

Although MPI does work on heterogeneous workstation clusters, its focus is on high-performance communication on large multi-processors\textsuperscript{43}. This results in a rich variety of communication mechanisms. However, the MPI API lacks dynamic resource management, which is necessary for fault tolerant applications.

In the following Sections, we introduce the main components of MPI. Furthermore, we briefly explain some MPI functions, which are used in Section 7.

4.1.1. Process Topology and Session Management

To tell the truth, there is no real session management in MPI. Each process of a MPI application is started independent from the others. At some point, the individual processes are exchanging messages, or are synchronized at a barrier. Finally, they shut-down, thus terminating the application. The distribution of the individual processes to the different processing entities (e.g. processors of a multi-processor) is handled by the underlying vendor library.

- \textbf{int MPI_Init(int *argc, char ***argv);} - initializes process for MPI.
- \textbf{int MPI_Finalize(void);} - releases process from MPI.

Furthermore, the user can specify the process topology within a group (see Section 4.1.2). Besides creating a convenient name space, the specification can be used by the runtime system to optimize communication along the physical interconnection between the nodes\textsuperscript{39}.

4.1.2. Grouping Mechanisms

A special feature of MPI is support for implementing parallel libraries. Many functions are provided to encapsulate communication within parallel libraries. These functions define a group scope for communication, synchronization, and other related operations of a library. This is done by introducing the concepts of communicators, contexts, and groups.

Communicators are the containers of all communication operations within MPI. They consist of participants (members of groups) and a communication context. Communication is either between members of one group (intra-communication), or between members of different groups (inter-communication). While the first kind of communication provides point-to-point communication and collective communication (such as broadcasts), the second kind only allows point-to-point communication. After initializing MPI for a process, two communicators are predefined. The MPI_COMM_WORLD communicator includes all processes which can communicate with the local process (including the local process). In contrast, the MPI_COMM_SELF communicator only includes the local process.

A group defines the participants of communication or synchronization operations. They define a unique order on their members, thus associating a rank (identifier of member within the group) to each member process. The predefined group MPI_GROUP_EMPTY defines an empty group.

The following functions provide information on a group or its members.

- \textbf{int MPI_Comm_size(MPI_Comm com, int* nprocess);} - returns the number of participating processes of communicator com.
- \textbf{int MPI_Comm_rank(MPI_Comm com, int* rank);} - returns rank of calling process.

A context defines the “universe” of a communicator. For intra-communicators, they guarantee that point-to-point communication does not interfere with collective communication. For inter-communicators, a context only insulates point-to-point communication, because collective operations are not defined.

4.1.3. Communication

There are two different communication methods. Group members can be either communicate pairwise, or they can communicate with all members of the group. The first method is called point-to-point communication, the second method is called collective communication. Furthermore, a communication operation can be blocking (it waits until the operation is done) or non-blocking (it does not wait).

Point-To-Point Communication

This class of communication operation defines communication between two processes. These processes can be either members of the same group (intra-communication), or they are members of two different groups (inter-communication). However, we only describe systems with one group (all processes). Therefore, we only use intra-communication.

\textcopyright{} The Eurographics Association and Blackwell Publishers 1998.
Usually, a message is attached to a message envelope. This envelope identifies the message and consists of the source or destination rank (process identifier), the message tag, and the communicator.

For blocking communication, the following functions are available:

- int MPI_Send(void *buf, int n, MPI_Datatype dt, int dest, int tg, MPI_Comm com); - sends the buffer buf, containing n items of datatype dt to process dest of communicator com. The message has the tag tg.
- int MPI_Recv(void *buf, int n, MPI_Datatype dt, int source, int tg, MPI_Comm com); - receives the message tagged with tg from process source of communicator com. The used buffer buf consist of n items of the datatype dt.

These functions are specifying the standard blocking communication mode, where MPI decides if the message is buffered. If the message is buffered by MPI, the send call returns without waiting for the receive post. If the message is not buffered, send waits until the message is successfully received by the respective receive call. Besides this standard mode, there are buffered, synchronous, and ready modes. More information on these modes can be found in the MPI specification papers39,40.

For non-blocking communication MPI_Isend and MPI_Irecv are provided for intermediate (I) communication. For buffered, synchronous, or ready communication modes, please refer to the MPI papers. After calling these functions, the buffers are send (or set while receiving). However, they should not be modified until the message is completely received.

- int MPI_Isend(void *buf, int n, MPI_Datatype dt, int dest, int tg, MPI_Comm com, MPI_Request* req); - sends the buffer buf, contain n items of datatype dt to process dest of communicator com. The message has the tag tg.
- int MPI_Irecv(void *buf, int n, MPI_Datatype dt, int source, int tg, MPI_Comm com, MPI_Request* req); - receives the message tagged with tg from process source of communicator com. The used buffer buf consist of n items of the datatype dt.

In addition to the blocking send and receive, the request handle req is returned. This handle is associated with a communication request object - which is allocated by these calls - and can be used to query this request using MPI_Wait.

- int MPI_Wait(MPI_Request* req, MPI_Status *stat); - waits until operation req is completed.

The last call we describe for point-to-point communication is MPI_Iprobe. This call checks incoming messages if they match the specified message envelope (source rank, message tag, communicator), without actually receiving the message.

- int MPI_Iprobe(int source, int tg, MPI_Comm com, int* flag, MPI_Status* stat); - checks incoming messages. The result of the query is stored in flag.

If flag is set true, the specified message is pending. If the specified message is not detected, flag is set to false. The source argument of MPI_Iprobe may be MPI_ANY_SOURCE, thus accepting messages from all processes. Similarly, the message tag can be specified as MPI_ANY_TAG. Depending on the result of MPI_Iprobe, receive buffers can be allocated and source ranks and message tags set.

Collective Communication

Collective Communication is only possible within a group. This implements a communication behavior between all members of the group, not only two members as in point-to-point communication.

We concentrate on two functions:

- int MPI_Bcast(void *buf, int n, MPI_Datatype dt, int root, MPI_Comm com); - broadcasts message buf of n items of datatype dt from root to all group members of communicator com, including itself.

While the first call synchronizes all processes of the group of communicator com, the second call broadcasts a message from group member root to all processes. A broadcast is received by the members of the group by calling MPI_Bcast with the same parameters as the broadcasting process, including root and com. Please note that collective operations should be executed in the same order in all processes. If this order between sending and receiving broadcasts is changed, a deadlock might occur. Similarly, the order of collective point-to-point operation should be the same too.

4.2 Parallel Virtual Machine - PVM

While MPI was designed for message-passing on multiprocessors, PVM was originally intended for message-passing within a heterogeneous network of workstations. In order to guarantee interoperability between independent computers, the concept of a virtual machine was introduced. While MPI supports only portability (a MPI-based application can be compiled on any system) but not interoperability, PVM processes can even communicate with processes build on completely different machines. Furthermore, processes can be started or terminated dynamically from a master process3, thus enabling dynamic resource management and fault tolerant applications.

³ Functions to start or terminate processes are integrated in MPI 2.0.
Generally, a parallel application using PVM3 is split into a master process and several slave processes. While the slaves do the actual work of the task, the master distributes data and sub-tasks to the individual slave processes. Finally, the master synchronizes with all slaves at a barrier, which marks the end of the parallel processing.

Before starting the parallel sessions, all designated machines of the cluster need to be announced in a hostfile. Furthermore, PVM demons must run on these machines. After running of the parallel sessions, all PVM demons (virtual machines) are shut down.

After this initialization, the master starts its execution by logging on to the running parallel virtual machine (PVM demon). Thereafter, it determines the available hardware configuration (number of available machines (nodes), ...), allocates the name space for the slaves, and starts these slaves by assigning a sub-task (program executable). After checking if all slaves are started properly, data is distributed (and sometimes collected) to the slaves.

At the end of the parallel computation, results are collected from the slaves. After a final synchronization at a common barrier, all slaves and the master log off from the virtual machine.

Next, we briefly introduce some commands for the process control. Furthermore, we introduce commands for distributing and receiving data. For details, please refer to the PVM book\(^\text{42}\).

### PVM Process Control

- `int pvm_mytid(void);` - logs process on to virtual machine.
- `int pvm_exit(void);` - logs process off from virtual machine.
- `int pvm_config(int* nproc, ...)` - determines number of available nodes (processes), data formats, and additional host information.
- `int pvm_spawn(char *task, ...)` - starts the executable task on a machine of the cluster.
- `int pvm_joingroup(char *groupname);` - calling process joins a group. All members of this group can synchronize at a barrier.
- `int pvm_lvgroup(char *groupname);` - leaving the specified group.
- `int pvm_barrier(char *groupname);` - wait for all group members at this barrier.
- `int pvm_kill(int tid)` - kill slave process with identifier tid.

### PVM Communication

- `int pvm_initsend(int opt)` - initializes sending of a message.
- `int pvm_pkint(int* data, int size, ..)` - encodes data of type int\(^\text{1}\) for sending.
- `int pvm_send(int tid, int tag, ..)` - sends data asynchronously (does not wait for an answer) to process tid with specified tag.
- `int pvm_bcast(char* group, int tag);` - broadcasts data asynchronously to all group members.
- `int pvm_mcast(int* tids, int n, int tag);` - broadcasts data synchronously to \(n\) processes listed in \(tids\).
- `int pvm_nrecv(int tid, int tag);` - non-blocking (does not wait if message has not arrived yet) receiving of message.
- `int pvm_recv(int tid, int tag);` - blocking receiving of message tag.
- `int pvm_upkint(int* data, int size, ..)` - decodes received data of type int.

There is only one active message buffer at a time. This determines the order of initialization, coding, and sending of the message.

### B. Pthread Programming

There are quite a number of thread models around, like the mthread package\(^\text{118}\) of the University of Erlangen-Nürnberg, the dots package\(^\text{12}\) of the University of Tübingen, the Compiler-Parallel-Support package of HP/Convex. There are NT-threads, Solaris-threads, and last but not least there is the IEEE POSIX thread standard (pthreads). In this tutorial, we will focus only on pthreads. Furthermore, all the examples are tested on SGI’s implementation of pthreads (available for IRIX 6.x and up).

The pthread standard defines an “Application Programming Interface” (API), as specified by POSIX standard 1003.1, or more specific: ISO/IEC 9945-1:1996 (ANSI/IEEE Std 1003.1, 1996 Edition). However, this standard does not define a particular implementation of this standard. Therefore, many definitions are opaque to the user, e.g. thread mapping, data types, etc...

The following text only gives a more or less brief introduction into pthread programming. Advanced features like real-time scheduling or attribute objects are only briefly mentioned or even completely ignored. For a more complete introduction into those topics, please refer to the books\(^\text{14, 87, 64, 57, 90}\) listed in Section 8.

### 4.3. Concurrency

There are some differences between programming of sequential programs and concurrent (parallel) programs. It is very important to realize that concurrent programs can behave completely differently, mainly because the notion of

\(^\text{1}\) There are commands for other data types, such as byte, double, as well.

a sequence is not really available on the process level, although it is available on thread the level.

First of all, the order of sequential programs is determined at all times. In parallel programs, however, it is not. There are no statements within the pthread standard which control the actual order pthreads are scheduled. Consequently, we cannot tell which pthread will be executed before the other pthread.

Second - critical sections. A sequential program does not need to make sure that data which is not completely changed might be already read in another part of the program, because a sequential program only performs one statement at a time. This is different with concurrent programs, where different threads might perform different statements at virtually the same time. Therefore, we need to protect those areas, which might cause inconsistent states, because the modifying thread is interrupted by a reading thread. These areas are called critical sections. The protection can be achieved by synchronizing the threads at the beginning of these critical sections.

Third - error handling. Another difference is error handling. While UNIX calls usually return an useful value, if execution was successful, a potential error code is returned to the general error variable errno. This is not possible using threads, because a second thread could overwrite the error code of a previous thread. Therefore, most pthread calls return directly an error code, which can be analyzed or printed onto the screen. Alternatively, the string library function

```c
char* strerror(int errno);
```

returns an explicit text string according to the parameter errno.

### 4.4. Controlling Pthreads

In this part, we discuss the life cycle of a pthread. The life cycle starts with the creation of the pthread, its work, and the end of its existence.

To start the life of a pthread, we need to execute the `pthread_create` command:

```c
int pthread_create( pthread_t *pthread_id, 
const pthread_attr_t* ptr, 
void* (*thread_routine) (void *), 
void *arg );
```

where
- `pthread_id` is the returned identifier of the created pthread.
- `pthread_attr_t` is the passed attribute structure. If NULL is passed, the default attributes are used.
- `thread_routine` is the name of the function which is called by the created pthread, and
- `arg` is a pointer to the parameter structures for this pthread.

If this function returns error code 0, it was successful. If an error was encountered, the return code specifies the encountered problem.

If a pthreads needs to know its identity, this identity can be established using the call

```c
pthread_t pthread_self(void);
```

where the pthread identifier of the current pthread is returned. However, the pthread identifier of another pthread is only known by its caller. If this information is not passed to the particular pthread, this pthread does not know the identifier of the other pthread.

Similar to the last call,

```c
int pthread_equal(pthread_t t1, 
 pthread_t t2);
```

determines if two pthread identifiers are referring to the same pthread. If `t1` is equal `t2` a nonzero value will be returned (“True”); if they are not equal, zero will be returned (“False”).

The life of a pthread usually terminates with a `pthread_exit` call. Although the pthread is terminated, the resources used by this pthread are still occupied, until the pthread is detached. Using the command

```c
int pthread_detach(pthread_t pthread_id);
```

explicitly detach a pthread, telling the operating system that it can reclaim the resources as soon as the pthread terminates.

If a pthread A needs to wait for termination of pthread B, the command

```c
int pthread_join(pthread_t pthreadB_id, 
 void **ret_val);
```

can be used. As soon as pthread B terminates, it joins pthread A, which is waiting at the `pthread_join` command. If pthread B is returning a result using the pointer `ret_val`, this pointer is accessible via `ret_val` of the `pthread_join` command. If `ret_val` is set to NULL, no return value will be available. `pthread_join` implicitly detaches the specified pthread.

An example for pthread creation can be found as listing 1 in Section 4.7.1.

### 4.5. Pthread Synchronization

One of the most important topics in thread programming is synchronization. Different resources (e.g. variables, fields, etc.) are shared by different threads. Therefore, the access to these resources needs to be protected. Usually, this protection for MUTual EXclusion is done by a mutex. However, other synchronization mechanisms are known, such as conditions and barriers.
4.5.1. Mutex Synchronization

A mutex protects a critical section in a program. Considering a scenario, where rendering information is stored in a special data structure - e.g. a FIFO queue - and two threads try to read information from that data structure, obviously, the access to this data structure is a critical section and the access must be limited to one thread at the time. Therefore, the data structure must be protected by a mutex.

**Initialization**

```c
pthread_mutex_t mutex = 
    PTHREAD_MUTEX_INITIALIZER;

int pthread_mutex_init(
    pthread_mutex_t *mutex,
    pthread_mutexattr_t *attr);

int pthread_mutex_destroy(
    pthread_mutex_t *mutex);
```

After memory allocation of the mutex structure, it must be initialized. For static allocation, we can simply assign the preprocessor macro `PTHREAD_MUTEX_INITIALIZER` to the mutex.

In most cases, however, we dynamically allocate a mutex. For these cases, we can use `pthread_mutex_init` to initialize the allocated mutex structure. The second parameter of this command is used to specify a mutex attribute object. This attribute object is not frequently used. Therefore, we pass NULL.

If no pthread is locking the mutex, we can destroy it using `pthread_mutex_destroy` before releasing the mutex structure memory. If the mutex is statically allocated and initialized, the explicit destruction of the mutex is not necessary.

**Using a Mutex**

```c
int pthread_mutex_lock(
    pthread_mutex_t *mutex);

int pthread_mutex_trylock(
    pthread_mutex_t *mutex);

int pthread_mutex_unlock(
    pthread_mutex_t *mutex);
```

Before entering a critical section in a parallel program, we need to lock the associated mutex using `pthread_mutex_lock`. If the mutex is already locked, the current pthread will be blocked, until the mutex is unlocked by the other pthread. The behavior if a pthread tries to lock a mutex which is already locked by the very same pthread is not defined. Either an error code will be returned, or this pthread will end up in a deadlock.

In case you do not want to wait on an already locked mutex, you can use `pthread_mutex_trylock`. This call returns EBUSY in case that the specified mutex is already locked by another pthread. At the end of a critical section you need to unlock the locked mutex using `pthread_mutex_unlock`.

An example for pthread mutexes can be found as listing 2 in Section 4.7.2.

**Semaphores**

Semaphores is a concept which is more or less a generalization of a mutex. While a mutex only is a binary representation of the state of a resource, a semaphore can be used as a counter ("counting semaphores"). Although the pthread standard does not specify semaphores, the POSIX semaphores can be used.

4.5.2. Condition Synchronization

While mutexes protect a critical section of a program, conditions are used to send messages on the state of shared data. Considering the classic user/producer problem, the producer signals a condition to the users that it has produced data which can be digested by the users.

Dave Butenhof\textsuperscript{14} says that condition variables are for signaling, not for mutual exclusion.

**Initializing**

```c
pthread_cond_t cond = 
    PTHREAD_COND_INITIALIZER;

int pthread_cond_init(
    pthread_cond_t *cond,
    pthread_condattr_t *condattr);

int pthread_cond_destroy(
    pthread_cond_t *cond);
```

Similarly to the mutex initialization, static and dynamic allocated condition structures need to be initialized using the respective commands. For our use, we always pass NULL to the `condattr` parameter. Further discussion of the attribute features can be found in Butenhofs book\textsuperscript{14}.

After use, the condition structures need to be destroyed before releasing the associated memory.

**Using conditions**

```c
int pthread_cond_wait(
    pthread_cond_t *cond,
    pthread_mutex_t *mutex);

int pthread_cond_timedwait(
    pthread_cond_t *cond,
    pthread_mutex_t *mutex,
    struct timespec *exp);
```

4.5.3. Barrier Synchronization

The last presented synchronization concept is the barrier synchronization. Unfortunately, this concept is not part of the current pthread standard (1996), but it is on the draft list for the next version.

Generally, a barrier synchronization stops threads at this barrier, until the specified number of threads arrive. Thereafter, all threads proceed. There are different suggestions how to implement barriers in the current pthread standard. We will present two examples of an implementation. The first one implements a barrier synchronization at the end of the life cycle of the threads by joining them in a cascade (see listing 2 in Section 4.7.2). However, this method is not suited for a barrier synchronization which is not at the end of the life cycle of the pthreads, but in the middle of the working program. In addition, it has some structural limitations, because each pthreads in the cascade needs to know its successor’s pthread identifier.

The second example is from Dave Butenhof book on POSIX threads\textsuperscript{14}. In this example, every pthread which waits at a barrier is decrementing the waiting pthread counter and checks if more pthreads are expected to wait at this barrier. If no further pthread is expected to wait, it broadcasts the other waiting pthreads that the appropriate number of pthreads arrived at the barrier. If the number of waiting pthreads is not reached, this pthreads starts waiting for the broadcast. This implementation of a barrier can be found as listing 4, Section 4.7.4.

4.6. Additional Topics

4.6.1. Concurrent Memory Visibility

As mentioned earlier, programming concurrent (parallel) systems is quite different from programming sequential systems. This is especially true for the view of the memory we are using within our parallel program.

Modern processors are buffering data into caches of different sizes and different levels. If more than one processor is working for one program, different caches are storing information. Therefore, the information visible by one processors (in its cache) might be not the same as visible to another processor (in its cache or the main memory). This problem becomes even worse if NUMA memory architectures are used, because checking for changes in different caches and different memory hierarchies is much more difficult.

The pthread standard defines situations when the memory view of the different threads (possibly running on different processors) is equal, providing that the memory has not changed after these commands.

- After starting pthreads (\texttt{pthread_create}), the started pthreads have the same memory view as their parent.
- After explicitly (\texttt{pthread_mutex_unlock}) or implicitly (conditions) unlocking mutexes, the pthreads
which are blocked at this mutex have the same memory view as the unlocking pthread.

- Furthermore, the memory view of terminated pthreads (canceled pthreads, exited pthreads, or simply returning from their thread function) is the same as of the pthread which joins the terminating pthreads.

- Finally, each pthread which is waked-up by a signaling or broadcasting pthread has the same memory view as the signaling or broadcasting pthread.

Apart from these situations, the same memory view can not be guaranteed. Although you might never encounter this problems on a particular system (it might be cache-coherent), you can never be sure.

### 4.6.2. Cancelation

```c
int pthread_cancel(pthread_t pthread_id);
int pthread_setcancelstate(int state, int* ostate);
int pthread_setcanceltype(int type, int* otype);
void pthread_testcancel(void);
```

Usually, a thread is executing a particular part of the program until the task is done and the thread is either returning to its parent thread (main thread), or exits. However, there are situations where the task of the thread becomes dispensable. In those cases, it is useful to cancel this thread.

In general, we need the pthread identifier of the pthread to be canceled. Without this identifier, we cannot cancel the pthread. To cancel a pthread, we call `pthread_cancel(pthread_id);`.

There are three different cancelation modes the user can choose from. First, there is the DISABLED mode, where the cancel state is set to PTHREAD_CANCEL_DISABLE (the value of the cancel type will be ignored). In this mode no cancelation is possible. It becomes meaningful to prevent data corruption, while the pthread is changing data. In this cases, the pthread disables cancelation until it has finished the modification. Thereafter, it enables cancelation again. Cancel requests issued while the cancelation is disabled, are queued until the cancelation state is enabled again.

If the cancelation state is set to PTHREAD_CANCEL_ENABLE, we can choose from two cancelation types: PTHREAD_CANCEL_DEFERRED (the default) or PTHREAD_CANCELASYNCHRONOUS. The second type indicates that the respective pthread should be canceled at any time from now. This might cause data corruption, deadlocks - pthreads which are locked at a mutex locked by the canceled pthread -, and so forth. This is really an emergency kind of cancelation. Better is the first cancelation type, which asks the pthread to stop at the next cancelation point. At implicit cancelation points like pthread_cond_wait, pthread_cond_timedwait, or pthread_join, the pthread cancel immediately after executing these commands. However, an explicit cancelation point can be set using pthread_testcancel. If a cancel request is pending, the pthread returns the value PTHREAD_CANCELED to a pthread which waits to join this pthread. If no cancel request is pending, the pthread_testcancel command immediately returns. Besides these implicit or explicit cancelation points, there are library calls or system calls which are implicit cancelation points. Generally, these calls can introduce some blocking behavior and are therefore good candidates for cancelation. Please refer to one of the pthread books for a list of these calls.

Please note, enabling cancelation is not a cancelation point. Therefore, you need to explicitly set a cancelation point after enabling cancelation.

Another feature of cancelation is the specification of an cleaning-up handler for the pthread to be canceled. This cleaning-up handler can close files, release memory, repair data modifications, and so forth. Please refer to Buthofs book[14] for more information on cleaning-up canceled pthreads.

### 4.6.3. Hints

In this Section, we provide some tips and hints on common problems and usage of pthreads on some systems.

#### Debugging

- **Thread races.** Never count on an execution order of pthreads. Generally, we can not assume a certain executing order of pthreads. The standard does not completely control the actual scheduling of the physical system. Furthermore, after creation of a pthread, you cannot count that this pthread will start before another pthread created after the first pthread.
- **Avoid potential deadlock situations.** Well, this sounds obvious. However, there are many unavoidable situations which are potential deadlock situations. If you use mutex hierarchies (lock one mutex after successfully locking a first mutex), you need to consider a back-off strategy in case that the second mutex locking will block the pthread, which keeps the first mutex.
- **Priority inversion.** If you use real-time priority scheduling (see Section 4.6.4), your scheduling strategy (FIFO) might schedule a pthread to run which tries to lock a mutex, locked by a pthread preempted by the first pthread. Mutual exclusion and scheduling performing a kind of contradictory execution which can cause a deadlock.
- **Sharing stacks.** Pthread attributes (Section 4.6.4) enable the user to share stack memory. If the size of this stack...
is too small for these pthreads, you will encounter some strange effects.

Performance

- **Mutexes** are not free. You should always carefully decide if you use a “big mutex” protecting one big piece of code, or a number of mutexes protecting more fine granularly critical sections.

- **IRIX Pthreads**. The current implementation of pthreads on SGI workstations maps the pthreads on sproc lightweight processes of the operating systems. Furthermore, the system decides if it starts additional sproc for an additional pthread. In my experience, this does not work very well. Therefore, you can tell the operating system that your pthreads are compute-bound, by setting the environment variable PT_ITC (setenv PT_ITC). This usually results in starting enough sprocs for all processors.

- **Solaris threads**. On Solaris 2.5, threads are not time-sliced. Therefore, we need to set the concurrency level to the number of started threads, in order to obtain a concurrent program execution. The respective command is thr_setconcurrency(<nthreads>);

I found a nice quote in Dave Butenhof’s book for those who are getting frustrated while debugging a concurrent program:

> Wisdom comes from experience, and experience comes from lack of wisdom.

### 4.6.4. Further topics

In this tutorial, we do not provide material on all pthread topics. In my experience, I have never needed features like one-time initialization, real-time scheduling, thread-specific data, thread attributes, and so forth. However, there are situations where you might need this features. A discussion of these additional features can be found in Butenhof’s book\(^4\).

### 4.7. Example Code

This part contains example code for pthread programming. Please note that we denote the thread which starts all other threads as main thread. Naturally, this thread is considered as a pthread too. Furthermore, we use the term pthread for all threads started by the main thread using the command pthread_create.

#### 4.7.1. Initializing Pthreads

The pthread program listed listing 1, starts five pthreads and passes a couple of values to the pthreads. Finally, the main pthread collects the pthread started first.

14, number of pthreads started (including the main thread).
until all started pthreads are joined. Please note that after the presented kind of barrier synchronization, no pthreads are running anymore.

131-137, the mutex is released.

4.7.3. Condition Example

The pthread program listed listing 3, starts two pthreads and passes a couple of values to the pthreads. Finally, the main pthread collects the started pthreads. The pthreads are alternating processing a shared variable using conditions to signal the state of the variable to the pthread.

18-24, type definition of shared data passed to the pthreads.

26, shared data definition.

39-72, allocates and initializes mutex (43-51) and conditions (53-70).

74-113, producer pthread. After waiting for two seconds (82) in order to make sure that consumer pthread waits at the condition, the producer locks the mutex (87-90), manipulates the shared data, setting the predicate to 0, marking that it has been processed by the producer (91-92), and signals to the consumer that the shared data is ready to process (93-96). Thereafter, the producer waits until the data is consumed by the consumer pthread (98-103). Each time the producer is waked up by a signal (99), it checks if the predicate is correct. If not, it continues waiting (98,103). This is to prevent wrong wake-ups of the waiting pthread. After waiting of the pthread at this condition, the mutex is unlocked (107-110). Please note that while waiting for the signal (99), the mutex is unlocked by the system and re-locked before returning to the user code.

115-151 consumer pthread. Similar to the producer pthread, the consumer locks the mutex (124-127) and waits at the condition for the proper signal (128-133). If a wrong signal is received which waked-up the consumer, the predicate is not set properly. Therefore, we continue waiting (128, 133). After successful receiving the proper signal, the consumer consumes the shared data (135), sets the predicate (136), and signals the consumption to the producer pthread (139-142). Thereafter, it unlocks the mutex (144-147) and continues waiting (124) for new data. Please note that the mutex is locked while producer/consumer are manipulating the shared data. The mutex is released by the pthreads, while they are waiting at the condition. If the mutex is unlocked while manipulating the data, a deadlock is usually the result, because the later signal might be received by the other pthread. Due to scheduling, it is possible that the pthread just waked-up by a wrong signal, therefore misses the correct signal.

159-166, producer and consumer are started. The producer/consumer cycles is performed 200 times. (Actually, only pointers to data structures should be passed. An integer does not always fit into the memory space of a pointer. 170-177, both pthreads are collected by the main thread. 179-193, resources are released.

4.7.4. Barrier Example

This Section describes the barrier example of the book by Dave Butenhof.

Three functions are defined; barrier_init and barrier_destroy define the initializing and destructor functions of the barrier. The function barrier_wait defines the entrance to the barrier. The pthreads at this barrier wait until a specified number of pthreads has arrived. This number is specified in barrier_init.

1-42 barrier header file barrier.h.

43-186 barrier code file barrier.c.

19-26 type definition of barrier.

72-88 barrier_init. This function initializes the barrier (72). The number of pthreads which need to wait at this barrier is specified with count (72,76). In 77, cycle is initialized. This variable is used to filter wrong wake-up signals. Finally, the barrier is made valid in 86.

93-125 barrier_destroy. This function removes the barrier (93). After checking if the barrier is valid (97), the barrier access mutex is locked (100). If any pthreads are still waiting at this barrier (108), this function is aborted (110). If no pthreads are waiting, the barrier is invalidized (113), the access mutex unlocked and released, and the condition is removed (122,123).

132-186 is the actual barrier function. The pthreads which enter this function are blocked (169) until the it receives a signal and the cycle has changed, since the pthread has entered this function. If the pthread which has just entered the barrier_wait function is the pthread all the other pthreads are waiting for, it changes the cycle (146), resets the counter (147), and broadcasts to all waiting pthreads that it has arrived (148).


Listing 1 - Initializing Pthreads

1 /*
2 * create.c  
3 * starting and terminating pthreads
4 */
5
6 #include <stdio.h>
7 #include <stdlib.h>
8 #include <limits.h>
9 #include <string.h>
10 #include <pthread.h>
11
12 #define NO_PTHREADS 6
13
typedef struct {
14     pthread_t main;
15     int pthread_no;
16     ident_t*;
17 } id,s;
18
19 void* thread_function(void* arg)
20 {
21     int i;
22     ident_t* info = (ident_t*) arg;
23     pthread_t self;
24
25     self = pthread_self();
26     for (i=0; i<NO_PTHREADS; i++)
27         if (pthread_equal(self,info->main)) {
28             printf(stderr,“Current pthread is main thread.\n”);
29             else
30                 printf(stderr,“Current pthread is thread \#%d.\n”,
31                     info->pthread_no);
32             pthread_exit((void*) &info->pthread_no);
33         }
34     return NULL;
35 }
36
37 int main(void)
38 {
39     int i,rc;
40     void *resp;
41     pthread_t ids[NO_PTHREADS];
42     ident_t *info=NO_PTHREADS+1);
43     info[0].main = pthread_self();
44     ids[i] = pthread_self();
45     info[0].pthread_no = 0;
46     info[i].main = pthread_self();
47     for (i=1; i<NO_PTHREADS; i++)
48         if (rc) {
49             printf(stderr,“ERROR - while creating pthread \#%e\n”,
50                     info[i].pthread_no, strerror(rc));
51             exit(-1);
52         } else
53             printf(stderr,“Main: Thread \#%d started.\n”,
54                     ids[i].pthread_no,
55                     info[i].pthread_no,
56                     pthread_detach(ids[i]);
57     for (i=0; i<NO_PTHREADS; i++)
58         if (rc) {
59             printf(stderr,“ERROR - while detaching pthread \#%e\n”,
60                     info[i].pthread_no, strerror(rc));
61             exit(-1);
62         } else
63             printf(stderr,“Joined pthread result is \#%e\n”,
64                     (int) *resp);
65         }
Listing 2 - Mutex Example

```c
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <pthread.h>

#define EDK 0
#define TRUE 1
#define FALSE 0

#define LOCK_TIMES 100000

typedef struct {
  int       i;
  int       rc;
  int       tries;
  pthread_t *mutex;
} param_t;

void* thread_function(void* arg)
{
  int     i;
  int     rc;
  int     tries;
  param_t *info = (param_t*) arg;
  for (i=0; i<100; i++) {
    if (info->tries == 0)
      read = TRUE;
    while (read)
      switch (rc) {
        case EDK:
          try++;
          info->mutextries++;
          break;
        case OK:
          for (j=0; j<LOCK_TIMES; j++)
            rc = pthread_mutex_unlock(info->mutex);
          if (rc)
            fprint(stderr,"ERROR - while unlocking mutex: %s\n",
                   strerror(rc));
          exit(-1);
          break;
        default:
          fprint(stderr,"ERROR - while trying to lock mutex: %s\n",
                 strerror(rc));
          sterror(rc);
          break;
      }
  } return NULL;
}

int main(void)
{
  int     i, rc;
  pthread_t *id[NP_THREADS+1];
  pthread_mutex_t *mutex;
  param_t  info[NP_THREADS+1];

  mutex = (pthread_mutex_t*) calloc(1, sizeof(pthread_mutex_t));
  if (mutex)
    fprint(stderr,"ERROR - while allocating mutex.\n");
  exit(-1);
}
```

Listing 3 - Condition Example

```c
1 /*
2 * cond.c
3 * pthread condition handling
4 *
5 */
6
7 #include <stdio.h>
8 #include <stdlib.h>
9 #include <unistd.h>
10 #include <string.h>
11 #include <pthread.h>
12
13 #define NOTHING -1
14 #define CREATED 0
15 #define MODIFIED 1
16 #define RES (void*) 0;
17
18 typedef struct {
19   int val;
20   pthread_mutex_t mutex;
21   pthread_cond_t *created;
22   pthread_cond_t *consumed;
23   int pred; /* shared predicate */
24 } buffer_t;
25
26 static buffer_t buffer;
27
28 void ferr(char *text, int rc)
29 {
30   fprintf(stderr, "%s: %s", text, strerror(rc));
31   exit(-1);
32 }
33
34 void message(char *text)
35 {
36   fprintf(stderr, "%s\n", text);
37 }
38
39 void init(void)
40 {
41   int rc;
42   buffer.mutex = (pthread_mutex_t*) malloc(sizeof(pthread_mutex_t));
43   if (!buffer.mutex)
44     ferr("ERROR - while allocating mutex",0);
45   if (rc)
46     ferr("ERROR - while init' mutex",rc);
47 }
48
49 rc = pthread_mutex_init(buffer.mutex, NULL);
50 if (rc)
51     ferr("ERROR - while init' mutex",rc);
52 }
53
54 buffer.created = (pthread_cond_t*) malloc(sizeof(pthread_cond_t));
55 if (!buffer.created)
56     ferr("ERROR - while allocating condition created",0);
57 }
58
59 buffer.consumed = (pthread_cond_t*) malloc(sizeof(pthread_cond_t));
60 if (!buffer.consumed)
61     ferr("ERROR - while allocating condition consumed",0);
62 
63 rc = pthread_cond_init(buffer.created, NULL);
64 if (rc)
65     ferr("ERROR - while init' condition created",rc);
66 }
67 rc = pthread_cond_init(buffer.consumed, NULL);
68 if (rc)
69     ferr("ERROR - while init' condition consumed",rc);
70 }
71 message("Classic producer/consumer problem is started ...");
72 
73 void producer(void *times)
74 {
75   int i,t,rc;
76   t = (int) times;
77   i = 1;
78
79   /* To slow down initial signal */
80   sleep(2);
81
82   while (i< t) {
83     rc = pthread_mutex_lock(buffer.mutex);
84     if (rc)
85       ferr("ERROR - while locking mutex",rc);
86     buffer.val = i+1;
87     buffer.pred = 0;
88     rc = pthread_cond_signal(buffer.created);
89     if (rc)
90       ferr("ERROR - while signaling created",rc);
91   }
92
93   do {
94     rc = pthread_cond_wait(buffer.consumed, buffer.mutex);
95     if (rc)
96       ferr("ERROR - while waiting on consumed",rc);
97     i++;
98     rc = pthread_mutex_lock(buffer.mutex);
99     if (rc)
100       ferr("ERROR - while locking mutex",rc);
101   } while (buffer.pred != -1);
102
103   fprintf(stderr, "Modified value No. %d= %d\n",i,buffer.val);
104   i++;
105
106   rc = pthread_mutex_unlock(buffer.mutex);
107   if (rc)
108     ferr("ERROR - while unlocking mutex",rc);
109 }
110
111 return NULL;
112 }
113
114 void *consumer(void *times)
115 {
116   int i,t,rc;
117   t = (int) times;
118   i = 1;
119
120   message("Consumer is started ...");
121 
122```
while (i<e) {
  rc = pthread_mutex_lock(buffer.mutex);
  if (rc) {
    ferr("ERROR - while locking mutex",rc);
  } do {
  rc = pthreadCond_wait(buffer.created,buffer.mutex);
  if (rc) {
    ferr("ERROR - while waiting on empty",rc);
  } while (buffer.pend != 0);
  buffer.val *=11;
  buffer.pend = 1;
  i++;
  rc = pthreadCond_signal(buffer.consumed);
  if (rc) {
    ferr("ERROR - while signaling consumed",rc);
  }
  rc = pthread_mutex_unlock(buffer.mutex);
  if (rc) {
    ferr("ERROR - while unlocking mutex",rc);
  }
  return NULL;
}
}
(void main(void)
{ pthread_t pthreadA, pthreadB;
  int  rc;
  init();
  rc = pthread_create(&pthreadA, NULL, producer, (void*) 2001);
  if (rc) {
    ferr("ERROR - while creating pthread",rc);
  }
  rc = pthread_create(&pthreadB, NULL, consumer, (void*) 2001);
  if (rc) {
    ferr("ERROR - while creating pthread",rc);
  }
  message("Main is waiting for pthreads ...");
  rc = pthread_join(pthreadA,NULL);
  if (rc) {
    ferr("ERROR - while joining pthread",rc);
  }
  rc = pthread_join(pthreadB,NULL);
  if (rc) {
    ferr("ERROR - while joining pthread",rc);
  }
  rc = pthread_mutex_destroy(buffer.mutex);
  if (rc) {
    ferr("ERROR - while destroying mutex",rc);
  }
  rc = pthread_cond_destroy(buffer.created);
  if (rc) {
    ferr("ERROR - while destroying condition created",rc);
  }
  rc = pthread_cond_destroy(buffer.consumed);
  if (rc) {
    ferr("ERROR - while destroying condition consumed",rc);
  }
  free(buffer.mutex);
  free(buffer.created);
  free(buffer.consumed);
  message("Done.");
}
Listing 4 - Barrier Example


```c
#include <pthread.h>
#include <errors.h>
#include "barrier.h"

typedef struct barrier_tag {
  pthread_mutex_t mutex; /* Control access to barrier */
  pthread_cond_t cv; /* wait for barrier */
  int valid; /* set when valid */
  int threshold; /* number of threads required */
  int counter; /* current number of threads */
  int cycle; /* alternate wait cycles (0 or 1) */
} barrier_t;

#define BARRIER_VALID 0xdcafe

#define barrier_init(barrier) { 
  pthread_mutex_init(&barrier->mutex, NULL);
  pthread_cond_init(&barrier->cv, NULL);
  barrier->valid = 1;
  barrier->counter = 0;
  barrier->threshold = 0;
  barrier->cycle = 1;
}

int barrier_init(barrier_t *barrier, int count) { 
  int status = pthread_mutex_init(&barrier->mutex, NULL);
  if (status != 0) return status;
  status = pthread_cond_init(&barrier->cv, NULL);
  if (status != 0) return status;
  pthread_mutex_destroy(&barrier->mutex);
  return status;
}

/*
 * Destroy a barrier when done using it.
 */

int barrier_destroy(barrier_t *barrier) { 
  int status, status2;
  if (barrier->valid != BARRIER_VALID) return EINVAL;
  pthread_mutex_destroy(&barrier->mutex);
  pthread_cond_destroy(&barrier->cv);
  returnEINVAL;
}

#define barrier_wait(barrier) { 
  int status;
  int status2 = pthread_cond_wait(&barrier->cv, &barrier->mutex);
  if (status2 != 0) return status2;
  status = pthread_mutex_lock(&barrier->mutex);
  if (status != 0) return status;
  barrier->counter ++;
  return EWOULDBLOCK;
}
```

/*
 * Wait for all members of a barrier to reach the barrier. When
 * the count (of remaining members) reaches 0, broadcast to wake
 * all threads waiting.
 */

int barrier_wait ( barrier_t *barrier )
{
  int status, cancel, temp, cycle;
  if ( (barrier->valid != BARRIER_VALID) )
    return EINVAL;
  status = pthread_mutex_lock ( &barrier->mutex );
  if ( (status != 0) )
    return status;
  cycle = barrier->cycle; /* Remember which cycle we're on */
  if ( (barrier->counter == 0) )
    barrier->cycle = barrier->cycle;
  barrier->counter = barrier->threshold;
  status = pthread_cond_broadcast ( &barrier->cv );
  /*
   * The last thread into the barrier will return status
   * -1 rather than 0, so that it can be used to perform
   * some special serial code following the barrier.
   */
  if ( (status != 0) )
    status = -1;
  } else {
    /*
      * Wait with cancellation disabled, because barrier_wait
      * should not be a cancellation point.
      */
    pthread_setcancelstate ( PTHREAD_CANCEL_DISABLE, &cancel );
    /*
     * Wait until the barrier's cycle changes, which means
     * that it has been broadcast, and we don't want to wait
     * anymore.
     */
    while ( (cycle == barrier->cycle) )
      status = pthread_cond_wait ( &barrier->cv, &barrier->mutex );
    if ( (status != 0) )
      break;
  }
  pthread_setcancelstate ( cancel, &cancel );
  /*
   * Ignore an error is unlocking. It shouldn't happen, and
   * reporting it here would be misleading -- the barrier wait
   * completed, after all, whereas returning, for example,
   * EINVAL would imply the wait had failed. The next attempt
   * to use the barrier will return an error, or hang, due
   * to whatever happened to the mutex.
   */
  pthread_mutex_unlock ( &barrier->mutex );
  return status; /* error, -1 for waker, or 0 */
}
PART TWO

Rendering

5. Parallel Polygonal Rendering

Many datasets in design, modeling, and scientific visualization are built from polygons and often from simple triangles. Frequently these datasets are very big (several millions to several tens of millions of triangles) as they describe a polygonal approximation to an underlying true surface. The size of these datasets often exceeds the processing and rendering capabilities of technical workstations. Parallel algorithms and parallel computers have often been seen as a solution for interactive rendering of such datasets.

Parallel rendering algorithms have been developed in different domains of computer graphics. Developers of graphics hardware have long recognized the need to partition the graphics pipeline amongst several processors in order to achieve fast rendering performance. These efforts resulted in highly specialized architectures that were optimized for particular algorithms and workloads.

As supercomputers became more powerful and less expensive it was a natural step to use them to render and display the results of the computations they were running. This had the advantage of saving time and bandwidth because the data did not need to be transferred from the supercomputer to a dedicated rendering engine. Rendering on supercomputers often does not constitute the most cost-effective solution, e.g., measured as dollars per rendering performance. However, there is no dedicated rendering hardware and all graphics algorithms are implemented in software, thus offering more flexibility in the choice of algorithms and supported rendering features.

This paper will describe and discuss different solutions to the problem of efficient rendering of polygonal models on parallel computers. We will start out with a description of the background of the problem, in particular the polygon rendering process, different rendering scenarios, and issues related to the architecture of the target platform. Then we will discuss ways to classify different parallel rendering algorithms which will provide insights into the properties of different strategies. Finally, we will describe different approaches to achieve load-balancing in parallel rendering systems.

For further study the reader is referred to the papers in the bibliography, in particular 124, and the proceedings of the Parallel Rendering Symposia and the Eurographics Rendering Workshops.

5.1. Background

5.1.1. Rendering Pipeline

In this paper we will only consider rendering of polygonal models using the standard rendering pipeline, i.e. we will not discuss ray-tracing or volume rendering. Figure 8 shows the principial steps in rendering of a polygonal model. The description of the model is stored on disk in some file format such as VRML. Before commencing the actual rendering process, the model must be loaded from disk into main memory and converted into an internal representation suitable for rendering. All further processing steps are then memory-to-memory operations. It should be noted that the order of primitives on disk and in the in-memory representation is arbitrary and is usually determined by the application. In particular, the order of primitives in the should not be relied upon when trying to load-balance parallel processors.

Geometry processing forms the first stage of the rendering pipeline. It includes the steps of transforming objects from their intrinsic coordinate system, e.g. model coordinates, into device coordinates, lighting, computation of texture coordinates, and clipping against the view frustum. Except for clipping, all operations in this stage are performed on vertex information. (Clipping operates on entire polygons which is, in particular on SIMD computers, often disrupting the data flow. The steps in the geometry pipeline can be rearranged such that clipping is postponed until the very end when vertices are reassembled into triangles for rasterization. Geometry processing needs mostly floating point operations to transform vertices and to support lighting calculations. Depending on the number of lights and the complexity of the lighting model geometry processing requires between several hundred and a few thousand floating point operations per vertex.

Rasterization converts primitives (typically triangles) described as screen-space vertices into pixels. The resulting pixels are then subjected to various fragment processing operations, such as texture mapping, z-buffering, alpha-blending etc. The final pixel values are written into the frame buffer from where they are scanned out onto the display. Most graphics systems implement rasterization and fra-
ment processing as a unit. One notable exception is the Pix-
eFlow system\textsuperscript{33}.

Rasterization and fragment processing are use predomi-
nantly fixed-point or integer computations. Depending on
the complexity of the fragment processing operations, be-
 tween 5 and up to 50 integer computations per pixel and
per triangle are required. Because rasterization is algorithmi-
cally simple yet requires such a huge number of operations
it is often implemented in hardware.

More details on the computational requirements for the
different stages in the rendering pipeline can be found for
instance in \textsuperscript{34} pp. 866–873.

Finally, the complete image is either sent to the screen for
display or written to disk. In many parallel rendering algo-
rithms this step forms a performance bottleneck as partial
images stored on different processors have to be merged in
one central location (the screen or a disk). (Although this
step should be included when measuring the end-to-end per-
formance of a parallel rendering system, some researchers
explicitly exclude this step due to shortcomings of their par-
ticular target platform\textsuperscript{28}.)

5.1.2. Single-frame vs. multi-frame rendering

Rendering polygon models can be driven by several needs.
If the model is only used once for the generation of a still
image, the entire rendering process outlined above has to
be performed. The creation of animation sequences requires
rendering of the same model for different values of time
and consequently varying values for time-dependent render-
ing parameters, e.g., view position, object location, or light
source intensities. Even though multi-frame rendering could
be handled as repeated single-frame rendering, it offers the
opportunity to exploit inter-frame coherence. For example,
access to the scene database can be amortized over sev-
eral frames and only the actual rendering steps (geometry
processing and rasterization) must be performed for every
frame. Other ways to take advantage of inter-frame coher-
ence will be discussed below.

5.1.2.1. Target Architectures

Parallel polygon rendering has been demonstrated on a large variety of platforms rang-
ing from small multi-processor system using off-the-shelf
microprocessors over multi-million dollar supercomputers
to graphics workstations built with special-purpose hard-
ware.

In spite of the many differences between those computers,
their rendering performance depends on a few key architec-
tural features:

- **Disk bandwidth** determines how fast the model can be
  loaded from file into memory. For off-line rendering, i.e.,
  storing the image on file instead of displaying it on screen,
  disk performance also affects how fast the final image can
  be written back. For large models, disk access time takes up
  an appreciable portion of the total rendering time and calls
  for high-performance disk subsystems like disk striping.

- **Inter-processor communication** is required both to ex-
  change or transfer model data between processors and to
  synchronize the operation of the parallel processors. The
  former calls for a connection providing high bandwidth for
  large data packages while the latter requires low latency for
  small data transfers. Often these two needs result in con-
  flicting technical requirements. The physical interconnection
  can be formed by a bus, shared memory, dedicated intercon-
  nection networks, or by a standard networking infrastruc-
  ture like Ethernet. A mismatch between rendering algorithm
  and communication infrastructure will lead to saturated net-
  works, low graphics performance, and underutilized CPUs.
  It should be noted that advertised peak bandwidth of a net-
  work technology is often based on raw hardware perfor-
  mance and may differ by as much as an order of magnitude
  from the bandwidth attainable by an application in practice.

- **Memory bandwidth** determines how fast the processor can
  operate on model and pixel information once that information
  has been loaded from disk or over the network. The effec-
  tive bandwidth is determined by the entire memory sub-
  system, including main memory and the various levels of
caching.

- **Compute power**. Both floating point and integer opera-
  tions must be matched to the algorithm. As note above, ge-
  ometry processing requires mostly floating point operations
  while rasterization uses mostly integer operations. The core
  rendering routines contain only few branches. Note that the
  available compute power affects only the computational por-
  tions of the algorithm. Often computation is outweighed by
  communication, which leads to the (often surprising and dis-
  appointing) effect that increases in compute power have little
effect on the overall rendering performance\textsuperscript{104}.

5.2. Algorithm Classification

For many years the classification of parallel rendering algo-
rithms and architectures has proven to be an elusive goal. We
will discuss several such classifications to gain some insight
into the design space and possible solutions.

5.2.1. Pipelining vs. Parallelism

Irrespective of the problem domain, parallization strategies
can be distinguished by how the problem is mapped onto the
parallel processors.

For pipelining the problem is decomposed into individual
steps that are mapped unto processors. Data travel through
the processors and are transformed by each stage in the
pipeline. For many problems, like the rendering pipeline
(sic!), such a partitioning is very natural. However, pipelining
usually offers only a limited amount of parallelism. Fur-
thermore, it is often difficult to achieve good load-balancing

\textcopyright{} The Eurographics Association and Blackwell Publishers 1998.
amongst the processors in the pipeline as the different functions in the pipeline vary in computational complexity.

To overcome such constraints pipelining is often augmented by replicating some or all pipeline stages. Data are distributed amongst those processors and worked on in parallel. If the algorithms executed by each of the processors are identical, the processors can perform their operation in lockstep, thus forming a SIMD (single-instruction, multiple-data) engine. If the algorithms contain too many data dependencies thus making SIMD operation inefficient, MIMD (multiple-instruction, multiple-data) architectures are more useful. SIMD implementations are usually more efficient as the processors can share instructions and require very little interprocessor communication or synchronization.

5.2.2. Object Partitioning vs. Image Partitioning

One of the earliest attempts at classifying partitioning strategies for parallel rendering algorithms took into consideration whether the data objects distributed amongst parallel processors belonged into object space, e.g. polygons, edges, or vertices, or into image space, i.e. collections of pixels such as portions of the screen, scanlines or individual pixels. Object-space partitioning is commonly used for the geometry processing portion of the rendering pipeline, as its operation is intrinsically based on objects. Most parallelization strategies for rasterizers employ image-space partitioning. A few architectures apply object-space partitioning in the rasterizer.

5.2.3. Sorting Classification

Based on the observation that rendering can be viewed as a sorting process of objects into pixels, different parallel rendering algorithms can be distinguished by where in the rendering pipeline the sorting occurs. Considering the two main steps in rendering, i.e. geometry processing and rasterization, there are three principal locations for the sorting step: Early during geometry processing (sort-first), between geometry processing and rasterization (sort-middle), and after rasterization (sort-last).

Figure 9 illustrates the three approaches. In the following discussion we will follow in referring to a pair of geometry processor and a rasterizer as a renderer.

Sort-middle architectures form the most natural implementation of the rendering pipeline. Many parallel rendering systems, both software and hardware, use this approach, e.g. 7, 31, 6, 25, 23, 125, 28. They assign primitives to geometry processors that implement the entire geometry pipeline. The transformed primitives are then sent to rasterizers that are each serving a portion of the entire screen. One drawback is the potential for poor load-balancing among the rasterizers due to uneven distribution of objects across the screen. Another problem of this approach is the redistribution of primitives after the geometry stage which requires a many-to-many communication between the processors. A hierarchical multi-step method to reduce the complexity of this global sort is described in 28.

Sort-last assigns primitives to renderers that generate a full-screen image of all assigned primitives. After all primitives have been processed, the resulting images are merged/composited into the final image. Since all processors handle all pixels this approach offers good load-balancing properties. However compositing the pixels of the partial images consumes large amounts of bandwidth and requires support by dedicated hardware, e.g. 33. Further, with sort-last implementations it is difficult to support anti-aliasing, as objects covering the same pixel may be handled by different processors and will only meet during the final compositing step. Possible solutions, like oversampling or A-buffers, increase the bandwidth requirements during the compositing step even further.

Sort-first architectures quickly determine for each primitive to which screen region(s) it will contribute. The primitive is then assigned to those renderers that are responsible for those screen regions. Currently, no actual rendering systems are based on this approach even though there are some

indicators that it may prove advantageous for large models and high-resolution images\textsuperscript{83}. In \textsuperscript{83}, Mueller claims that sort-first has to redistribute fewer objects between frames than sort-middle. Similar to sort-middle, it is prone to suffer from load-imbalance unless the workload is levelled using an adaptive scheme that resizes the screen regions each processor is responsible for.

5.2.4. Load Balancing

As with the parallel implementation of any algorithm the performance depends critically on balancing the load between the parallel processors. There are workload-related and design-related factors affecting the load balancing. We will first discuss workload issues and then describe various approaches to design for good load-balance.

5.2.4.1. Workload Characterization

Several properties of the model are important for analyzing performance and load-balancing of a given parallel rendering architecture. Clipping and object tesselation affect load-balancing during geometry processing, while spatial object distribution and primitive size mostly affect the balance amongst parallel rasterizers.

Clipping. Objects clipped by the screen boundaries incur more work than objects that are trivially accepted or rejected. It is difficult to predict whether an object will be clipped and load-imbalance can result as a consequence of one processor receiving a disproportionate number of objects requiring clipping. There are techniques that can reduce the number of objects that require clipping by enabling rasterizers to deal with objects outside of the view frustum\textsuperscript{94}.\textsuperscript{29}. This reduces the adverse affects of clipping on load-balancing to negligible amounts.

Tesselation. Some rendering APIs use higher order primitives, like NURBS, that are tesselated by the rendering subsystem. The degree of tesselation, i.e. the number of triangles per object, determines the amount of data expansion occurring during the rendering process. The degree of tesselation is often view-dependent and hence hard to predict a priori. The variable degree of tesselation leads to load imbalances as one processor’s objects may expand into more primitives than objects handled by another processor. Tesselation also affects how many objects need to be considered during the sorting step. In sort-first architectures, primitives are sorted before the tesselation, thus saving communication bandwidth compared to sort-middle architectures.

Primitive distribution. In systems using image-space partitioning, the spatial distribution of objects across the screen decides how many objects must be processed by each processor. Usually, objects are not distributed uniformly, e.g. more objects may be located in the center of the screen than along the periphery. This creates potential imbalances in the amount of work assigned to each processor. Below we will discuss different approaches to deal with this problem.

5.2.4.2. Designing for Load-Balancing

Several design techniques are used to compensate for load-imbalance incurred by different workloads. They can be distinguished as static, dynamic and adaptive.

On-demand assignment is a dynamic method that relies on the fact that there are many more tasks (objects or pixels) than there are processors. New work is assigned to the first available, idle processor. Except during initialization and for the last few tasks, every processor will be busy all the time. The maximum load imbalance is bounded by the difference in processing time between the smallest (shortest processing time) and largest (longest processing time) task. The ratio of the number of tasks and the number of processors is called the granularity ratio. Selecting the granularity ratio requires a compromise between good load balancing (high granularity ratio) and overhead for instance due to large overlap factor (low granularity ratio). The optimal granularity ratio depends on the model, typical values range from about 4 to 32. Care must be taken when applying this technique to geometry processing: Some graphics APIs (like OpenGL) require that operations are performed in the exact order in which they were specified, e.g. objects are not allowed to “pass each other” on their way through the pipeline. MIMD geometry engines using on-demand assignment of objects could violate that assumption and must therefore take special steps to ensure temporal ordering, e.g. by labeling objects with time stamps.

Interleaving is a static technique which is frequently used in rasterizers to decrease the sensitivity to uneven spatial object distributions. In general, the screen is subdivided into regions, e.g. pixels, scanlines, sets of scanlines, sets of pixel columns, or rectangular blocks. The shape and the size of these regions determines the overlap factor. For a given region size, square regions minimize the overlap factor\textsuperscript{37}. Among n processor, each processor is responsible for every n-th screen region. The value n is known as the interleave factor. Since clustering of objects usually occurs in larger screen regions and since every object typically covers several pixels, this technique will eliminate most load-imbalance stemming from non-uniform distribution of objects. Interleaving makes it harder to exploit spatial coherence as neighboring pixels (or scanlines) are assigned to different processors. Therefore, the interleave factor, i.e. the distance be-
between pixels/scanlines assigned to the same processor, must be chosen carefully. Several groups have explored various aspects of interleaving for parallel rasterization, e.g.\textsuperscript{52}.

Adaptive scheduling tries to achieve balanced loading of all processors by assigning different number of tasks depending on task size. For geometry processing this might mean to assign fewer objects to processors that are receiving objects that will be tessellated very finely. In image-space schemes this means that processors are assigned smaller pixel sets in regions with many objects, thus equalizing the number of objects assigned to each processor.

Adaptive scheduling can be performed either dynamically or statically. Dynamic adaptation is achieved by monitoring the load-balance and if necessary splitting tasks to off-load or statically. Dynamic adaptation is achieved by monitoring the number of objects assigned to each processor. Dynamic schemes this means that processors are assigned smaller objects that will be tesselated very finely. In image-space schemes this means that processors are assigned smaller regions that will be tesselated very finely. In either case, the processors are working on independent frames and there is no communication between processors. Determining the balance between polygon transformation (generation) and polygon rasterization (consuming) is not suitable for interactive rendering sequence. It is not suitable for interactive rendering as it typically introduces large latencies between the time a frame is issued by the application and when it appears on the screen.

The application programming interface (API) impacts how efficiently the strategies outlined above can be implemented. Immediate-mode APIs like OpenGL or Direct3D do not have access to the entire model and hence do not allow global optimizations. Retained-mode APIs like Phigs, Performer, OpenGL Optimizer, Java3D and Fahrenheit maintain an internal representation of the entire model which supports partitioning of the model for load-balancing.

5.2.4.3. Data Distribution and Scheduling In distributed memory architectures, e.g. clusters of workstations or message-passing computers, object data must be sent explicitly to the processors. For small data sets, one can simply send the full data set to every processor and each processor is then instructed which objects to use. This approach fails however for large models either because there is not enough storage to replicate the model at every processor and/or the time to transfer the model is prohibitive due the bandwidth limitations of the network.

Therefore, most implementations replicate only small data structures like graphics state, e.g. current transformation matrices, light source data, etc., and distribute the storage for large data structures, primarily the object descriptions and the frame buffer.

For system using static assignment of rendering tasks object data have to be distributed only during the initialization phase of the algorithm. This makes it easy to partition the algorithm into separate phases that can be scheduled consecutively.

For dynamic schemes data must be distributed during the entire process. Therefore processors cannot continuously work rendering objects but must instead divide their available cycles between rendering and communicating with other processors. Such an implementation is described in\textsuperscript{23}: The system implements a sort-middle architecture where each processor works concurrently on geometry processing and rasterization, i.e. producing and consuming polygons. The advantage is that only a small amount of memory must be allocated for polygons to be transferred between processors. Determining the balance between polygon transformation (generation) and polygon rasterization (consuming) is not obvious. However,\textsuperscript{23} states that the overall system performance is fairly insensitive to that choice.

5.2.5. Summary

Parallel rendering of polygonal datasets faces several challenges most importantly load-balancing. Polygon rendering proceeds in two main steps: geometry processing and rasterization. Both steps have unique computational and communication requirements.

For geometry processing load balancing is usually achieved using on-demand assignment of objects to idle processors. For rasterization, interleaving of pixels or scanlines mostly eliminates load-balancing problems at the expense of less inter-pixel or inter-scanline coherence for each processor. Adaptive load-balancing schemes estimate or measure the workload and divide the screen into regions that will create approximately equal workload.

6. Parallel Volume Rendering

6.1. Introduction

Volume rendering\textsuperscript{56} is a powerful computer graphics technique for the visualization of large quantities of 3D data. It is
specially well suited for three dimensional scalar \cite{61, 27, 119, 100} and vector fields \cite{21, 74}. Fundamentally, it works by mapping quantities in the dataset (such as color, transparency) to properties of a cloud-like material. Images are generated by modelling the interaction of light with the cloudy materials \cite{129, 76, 75}. Because of the type of data being rendered and the complexity of the lighting models, the accuracy of the volume representation and of the calculation of the volume rendering integrals \cite{11, 55, 54} are of major concern and have received considerable interest from researchers in the field.

A popular alternative method to (direct) volume rendering is isosurface extraction, where given a certain value of interest \( \lambda \in \mathbb{R} \), and some scalar function \( f : \mathbb{R}^3 \rightarrow \mathbb{R} \), a polygonal representation for the implicit surface \( f(x, y, z) = \lambda \) is generated. There are several methods to generate isosurfaces \cite{66, 77, 84, 89}, the most popular being the marching cubes method \cite{66}. Isosurfaces have a clear advantage over volume rendering when it comes to interactivity. Once the models have been polygonized (and simplified \cite{107} – marching cubes usually generate lots of redundant triangles), hardware supported graphics workstation can be used to speed up the rendering. Isosurfaces have several disadvantages, such as lack of fine detail and flexibility during rendering (specially for handling multiple transparent surfaces), and its binary decision process where surfaces are either inside or outside a given voxel tends to create artifacts in the data (there is also an ambiguity problem, that has been addressed by later papers like \cite{89}).

### 6.1.1. Volumetric Data

Volumetric data comes in a variety of formats, the most common being (we are using the taxonomy introduced in \cite{117}) cartesian or regular data. Cartesian data is typically a 3D matrix composed of voxels (a voxel can be defined in two different ways, either as the datum in the intersection of each three coordinate aligned lines, or as the small cube, either definition is correct as long as used consistently), while the regular data has the same representation but can also have a scaling matrix associated with it.

Irregular data comes in a large variety, including curvilinear data, that is data defined in a warped regular grid, or in general, one can be given scattered (or unstructured) data, where no explicitly connectivity is defined. In general, scattered data can be composed of tetrahedra, hexahedra, prisms, etc. An important special case is tetrahedral grids. They have several advantages, including easy interpolation, simple representation (specially for connectivity information), and the fact that any other grid can be interpolated to a tetrahedral one (with the possible introduction of Steiner points). Among their disadvantages is the fact that the size of the datasets tend to grow as cells are decomposed into tetrahedra. In the case of curvilinear grids, an accurate decomposition will make the cell complex contain five times as many cells. More details on irregular grids are postponed until Section 6.3.

### 6.1.2. Interpolation Issues

In order to generate the cloud-like properties from the volumetric data, one has to make some assumptions about the underlying data. This is necessary because the rendering methods typically assume the ability to compute values as a continuous function, and (for methods that use normal-based shading) at times, even derivatives of such functions anywhere in space. On the other hand, data is given only at discrete locations in space usually with no explicit derivatives. In order to correctly interpolate the data, for the case of regular sampled data, it is generally assumed the original data has been sampled at a high enough frequency (or has been low-pass filtered) to avoid aliasing artifacts \cite{47}. Several interpolation filters can be used, the most common by far is to compute the value of a function \( f(x, y, z) \) by trilinearly interpolating the eight closest points. Higher order interpolation methods have also been studied \cite{16, 72}, but the computational cost is too high for pratical use.

In the case of irregular grids, the interpolation is more complicated. Even finding the cell that contains the sample point is not as simple or efficient as in the regular case \cite{85, 96}. Also, interpolation becomes much more complicated for cells that are not tetrahedra (for tetrahedra a single linear function can be made to fit on the four vertices). For curvilinear grids, trilinear interpolation becomes dependant on the underlying coordinate frame and even on the cell orientation \cite{127, 45}. Wilhelms et al. \cite{127} proposes using inverse distance weighted interpolation as a solution to this problem. Another solution would be to use higher order interpolation. In general, it is wise to ask the creator of the dataset for a suitable fitting function.

### 6.1.3. Optical Models for Volume Rendering

Volume rendering works by modelling volume as cloud cells composed of semi-transparent material which emits its own light, partially transmits light from other cells and absorbs some incoming light \cite{128, 73, 75}. Because of the importance of a clear understanding of such a model to rendering both, regular and irregular grids, the actual inner workings of one such mechanism is studied here. Our discussion closely follows the one in \cite{128}.

We assume each volume cells (differentially) emits light of a certain color \( E(\lambda)(x, y, z) \), for each color channel \( \lambda \) (red, green and blue), and absorbs some light that comes from behind (we are assuming no multiple scattering of light by particles – our model is the simplest “useful” model – for a more complete treatment see \cite{73}).

Correctly defining opacity for cells of general size is slightly tricky. We define the differential opacity at some depth \( z \) to be \( \Omega(z) \). Computing \( T(z) \), the fraction of light transmitted through depth 0 to \( z \) (assuming no emission of light inside the material), is simple, we just need to notice that the amount of transmitted light at \( z + \Delta z \) is just the
amount of light at $z$ minus the attenuation $\Omega(z)$ over a distance of $\Delta z$:

$$T(z + \Delta z) = T(z) - \Omega(z)T(z)\Delta z$$

(1)

what (after making a division by $\Delta z$ and taking limits) implies

$$\frac{dT(z + \Delta z)}{d\Delta z} = -\Omega(z)T(z)$$

(2)

The solution to this linear equation of the first order 19 with boundary condition $T(0) = 1$ is:

$$T(z) = e^{\int_{0}^{z} \Omega(u)du}$$

(3)

The accumulated opacity over a ray from front-to-back inside a cell of depth $d$ is $(1 - T(d))$. An important special case is when the cell has constant differential opacity $\Omega$, in this case $T(z) = e^{-\Omega z}$. Before we continue, we can now solve the question of defining differential opacity $\Omega$ from the unity opacity (usually user defined and saved in a transfer function table). A simple formula can express $\Omega$ in terms of $O$:

$$\Omega = \log\left(\frac{1}{1-O}\right)$$

(4)

If the model allows for the emission of light inside the material, a similar calculation can be used to calculate the intensity $I_\lambda$ for each color channel inside a cell. In this case using an initial intensity $I_\lambda(0) = 0$, the final system and solutions are as follows:

$$\frac{dI_\lambda(z)}{dz} = -\Omega(z)I_\lambda(z) + E_\lambda(z)$$

(5)

$$I_\lambda(z) = T(z) \int_{0}^{z} \frac{E_\lambda(v)}{T(v)}dv$$

(6)

Specializing the solution for constant color and opacity cells (as done above) we get the simple solution:

$$I_\lambda(z) = \frac{E}{\Omega}(1 - e^{-\Omega z})$$

(7)

Usually, for computational efficiency, the exponential in the previous equation is approximated by its first terms in the Taylor series. 128, 73, 75 describe in detail analytical solutions under different assumptions about the behavior of the opacity and emitted colors inside the cells, extensions to more complex light behavior and the several tradeoffs of approximating the exponentials with linear functions.

The previous equations show how to calculate the continuous color and opacity intensity, usually this calculation is done once for every cell, and the results from each cell are composited in a later step. Compositing operators were first introduced in 95, and are widely used. The most used operator in volume visualization is the over operator, its operation is basically to add the brightness of the current cell to the attenuated brightness of the one behind, and in the case of front-to-back compositing update the opacities of the cells. The equations for the over operator are:

$$C_o = C_o + C_b(1 - O_a)$$

(8)

$$O_o = O_a + O_b(1 - O_a)$$

(9)

It is important to note, that in these equations the colors are saved pre-multiplied by the opacities (i.e., the actual color is $C_o/O_o$), this saves one multiplication per compositing operation.

### 6.1.4. Ray Tracing

A popular method to generate images from volume data is to use ray tracing or ray casting 146, 61. Ray casting works by casting (at least) one ray per image pixel into volume space, point sampling the scene with some lighting model (like the one just presented) and compositing the samples as described in the previous section. This method is very flexible and extremely easy to implement. There are several extensions of basic ray casting to include higher order illumination effects, like discrete ray tracing 130, and volumetric ray tracing 116. Both of these techniques take into account global illumination effects incorporating more accurate approximations of the more general rendering equation 84.

Because of its size, volumetric ray casting (and ray tracing) is very expensive. Several optimizations have been applied to ray tracing 62, 63, 24. One of the most effective optimizations are the presence acceleration techniques, that exploit the fact volumetric data is relatively sparse 62, 63, 24, 132, 131, Levoy 62 introduced the idea by using an octree 101 to skip over empty space. His idea was further optimized by Danskin and Hanrahan 24 to not only skip over empty space, but also to speed up sampling calculations over uniform regions of the volume. Another important acceleration techniques include adaptive image sampling and early ray termination. Recently, Lacroute and Levoy 59 have introduced a hybrid method that combines some of the previous optimizations in a very efficient class of volume rendering algorithms.
PARC – Hardware-Based Presence Acceleration

Avila, Sobierajski and Kaufman\textsuperscript{9,115} introduced the idea of exploiting the graphics hardware on workstations to speed up volume rendering. First, they introduce PARC (Polygon Assisted Ray Casting)\textsuperscript{9}, a technique that uses the Z-buffer\textsuperscript{30} to find the closest and farthest possibly contributing cells. Later, a revised technique\textsuperscript{115} is proposed that (still using the Z-buffer) can produce a better approximation of the set of contributing cells.

Their algorithm consists of first creating a polygonal representation of the set of contributing cells (based on axis aligned quadrilaterals) from a coarse volume (see Figure 10). The coarse volume is calculated by grouping neighboring voxels together, creating supervoxels. Each supervoxel is then tested for the presence of interesting voxels (i.e., voxels that belong to the range of voxels mapped to non-zero intensities and opacities by the transfer functions). All six external faces of supervoxels are then marked based on its possible visibility (the second method seems to need to project all the faces).

\begin{figure}[ht]
\centering
\includegraphics[width=0.5\textwidth]{fig10.png}
\caption{Polygon Assisted Ray Casting.}
\end{figure}

In order to perform the actual rendering, in the first method (called Depth Buffer PARC), all the visible quadrilaterals are transformed and scan-converted twice. Once for finding the first non-empty front voxel, and again to determine the final integration location. In the second method (called Color Buffer PARC), a sweep along the closest major axis is generated by coloring the PARC cubes with power of two numbers (so they do not interfere with each other), what leaves a footprint of the intervals \((t_i, t_i+1]\) that can be used to better sample the regions having interesting voxels. This can be quite a savings, given that volumes are quite sparse (most of the time, only 5-10\% of a volume contains any lighting and shading information for a given set of transfer functions).

6.1.5. Splatting or Projection

Ray casting, described in Section 6.1.4, works from the image space to the object space (volume dataset). Another way of achieving volume rendering is to reconstruct the image from the object space to the image space, by computing for every element in the dataset its contribution to the image. Several such techniques have been developed\textsuperscript{27,122}.

Westover’s PhD dissertation describes the Splatting technique. In splatting, the final image is generated by computing for each voxel in the volume dataset its contribution to the final image. The algorithm works by virtually “throwing” the voxels onto the image plane. In this process every voxel in the object space leaves a footprint in the image space that will represent the object. The computation is processed by virtually “peeling” the object space in slices, and by accumulating the result in the image plane.

Formally the process consists of reconstructing the signal that represents the original object, sampling it and computing the image from the resampled signal. This reconstruction is done in steps, one voxel at a time. For each voxel, the algorithm calculates its contribution to the final image, its footprint, and then it accumulates that footprint in the image plane buffer. The computation can take place in back-to-front or front-to-back order. The footprint is in fact the reconstruction kernel and its computation is key to the accuracy of the algorithm. Westover\textsuperscript{122} proves that the footprint does not depend on the spatial position of voxel itself (for parallel projections), thus he is able to use a lookup table to approximate the footprint. During computation the algorithm just need to multiply the footprint with the color of the voxel, instead of having to perform a more expensive operation.

Although projection methods have been used for both regular and irregular grids, they are more popular for irregular grids. In this case, projection can be sped up by using the graphics hardware (Z-buffer and texture mapping)\textsuperscript{109}.

6.2. Parallel Volume Rendering of Regular Grids

Here, we present a high performance parallel volume rendering engine for our PVR system. Our research has introduced two contributions to parallel volume rendering:

\begin{itemize}
\item content-based load balancing
\item pipelined compositing
\end{itemize}

Content-based load balancing (Section 6.2.2) introduces a method to achieve better load balancing in distributed memory MIMD machines. Pipelined compositing (Section 6.2.3) proposes a component dataflow for implementing the Parallel Ray Casting pipeline.

The major goal of the research presented is to develop algorithms and code for volume rendering extremely large datasets at reasonable speed with an aim on achieving real-time rendering on the next generation of high-performance parallel hardware. The sizes of volumetric data we are primarily interested are in the approximate range of 512-by-512-by-512 to 2048-by-2048-by-2048 voxels. Our primary hardware focus is on distributed-memory MIMD machines, such as the Intel Paragon and the Thinking Machines CM-5.

\footnotesize{© The Eurographics Association and Blackwell Publishers 1998.}
A large number of parallel algorithms for volume rendering have been proposed. Schroeder and Salem have proposed a shear based technique for the CM-2 that could render 128³ volumes at multiple frames a second, using a low quality filter. The main drawback of their technique is low image quality. Their algorithm had to redistribute and remap the dataset for each view change. Montani et al.

61 developed a distributed memory ray tracer for the nCUBE, that used a hybrid image-based load balancing and context sensitive volume distribution. An interesting feature of their algorithm is the use of clusters to generate higher drawing rates at the expense of data replication. However, their rendering times are well over interactive times. Using a different volume distribution strategy but still a static data distribution, Ma et al.

67 have achieved better frame rates on a CM-5. In their approach the dataset is distributed in a K-d tree fashion and the compositing is done in a tree structure. Others 61, 15, 16 have used similar load balancing schemes using static data distribution, for either image compositing or ray dataflow compositing. Nieh and Levoy

88 have parallelized an efficient volume ray caster cite:Levoy:1990:ERT and achieved very impressive performance on a shared memory DASH machine.

6.2.1. Performance Considerations

In analyzing the performance of parallel algorithms, there are many considerations related to the machine limitations, like for instance, communication network latency and throughput. Latency can be measured as the time it takes a message to leave the source processor and be received at the destination end. Throughput is the amount of data that can be sent on the connection per unit time. These numbers are particularly important for algorithms in distributed memory architectures. They can change the behavior of a given algorithm enough to make it completely impractical.

Throughput is not a big issue for methods based on volume ray casting that perform static data distribution with ray dataflow as most of the communication is amortized over time 71, 15. On the other hand, methods that perform compositing at the end of rendering or that have communication scheduled as an implicit synchronization phase have a higher chance of experiencing throughput problems. The reason for this is that communication is scheduled all at the same time, usually exceeding the machines architectural limits. One should try to avoid synchronized phases as much as possible.

Latency is always a major concern, any algorithm that requires communication pays a price for using the network. The start up time for message communication is usually long compared to CPU speeds. For instance, in the iPSC/860 it takes at least 200μs to complete a round trip message between two processors. Latency hiding is an important issue in most algorithms, if an algorithm often blocks waiting for data on other processors to continue its execution, it is very likely this algorithm will perform badly. The classic ways to hide latency is to use pipelining or pre-fetching.

Even though latency and throughput are very important issues in the design and implementation of a parallel algorithm, the most important issue by far is load balancing. No parallel algorithm can perform well without a good load balancing scheme.

Again, it is extremely important that the algorithm has as few inherently sequential parts as possible if at all. Amadahl’s law shows how speed up depends on the parallelism available in your particular algorithm and that any, however small, sequential part will eventually limit the speed up of your algorithm.

Given all the constraints above, it is clear that to obtain good load balancing one wants an algorithm that:

• Needs low throughput and spreads communication well over the course of execution.
• Hides the latency, possibly by pipelining the operations and working on more than one image over time.
• Never causes processors to idle and/or wait for others without doing useful work.

A subtle point in our requirements is in the last phrase, how do we classify useful work? We define useful work as the number of instructions Iop executed by the best sequential algorithm available to volume render a dataset. Thus, when a given parallel implementation uses a suboptimal algorithm, it ends up using a much larger number of instructions than theoretically necessary as each processor executes more instructions than Iop (P denotes the number of processors). Clearly, one needs to compare with the best sequential algorithm as this is the actual speed up the user gets by using the parallel algorithm instead of the sequential one.

The last point on useful work is usually neglected in papers on parallel volume rendering and we believe this is a serious flaw in some previous approaches to the problem. In particular, it is widely known that given a transfer function and some segmentation bounds, the amount of useful information in a volume is only a fraction of its total size. Based on this fact, we can claim that algorithms that use static data distribution based only on spatial considerations are presenting “efficiency” numbers that can be inaccurate, maybe by a large margin.

To avoid the pitfalls of normal static data distribution, we present in the next section a new way to achieve realistic load balancing. Our load balancing scheme, does not scale linearly, but achieves very fast rendering times while minimizing the “work” done by the processors.

6.2.2. Content-Based Load Balancing

This section explains our approach to load balancing, which is able to achieve accurate load balancing even when using presence acceleration optimizations. The original idea
of our load balancing technique came from the PARC 9 acceleration technique. We notice that the amount of "work" performed by a presence accelerated ray caster is roughly directly proportional to the number of full supervoxels contained in the volume.

We use the number of full supervoxels a given processor is assigned as the measure of how much work is performed by that particular processor. Let \( P \) denote the number of processors, and \( c_i \) the number of full supervoxels processor \( i \) has. In order to achieve a good load balancing (by our metric) we need a scheme that minimizes the following function for a partition \( X = (c_1, c_2, \ldots) \):

\[
\frac{1}{n} \sum_{i \neq j} |c_i - c_j|, \quad \forall i, j \leq P
\]

Equation 10 is very general and applies to any partition of the dataset \( D \) into disjoint pieces \( D_i \). In our work we have tried to solve this optimization problem in a very restricted context. We have assumed that each \( D_i \) is convex. (We show later that this assumption makes it possible to create a fixed depth sorting network for the partial rays independently calculated each disjoint region.) Furthermore, we only work with two very simple subdivisions: slabs and a special case of a BSP-tree.

Before we go any further, it is interesting to study the behavior of our load balancing scheme in the very simple case of a slab subdivision of the volume \( D \). Slabs (see Figure 11) are consecutive slices of the dataset aligned on two major axes. Given a volume \( D \), with \( s \) superslices and \( p \) processors with the restriction that each processor gets contiguous slices, the problem of calculating the "best" load balancing partition for \( p \) processors consists of enumerating all the \( (s-1)(s-2) \ldots (s-p+1) \) ways of partitioning \( D \), and choosing the one that minimizes Equation 10.

![Figure 11: During slab-based load balancing, each processor gets a range of continuous data set slabs. The number of full supervoxels determines the exact partition ratio.](image)

The problem of computing the optimal (as defined by our heuristic choice) load balance partition indices can be solved naively as follows. We can compute all the possible partitions of the integer \( n \), where \( n \) is the number of slabs, into \( P \) numbers, where \( P \) is the number of processors (it is actually a bit different, as we need to consider addition non-associative). For example, if \( n = 5 \), and \( P = 3 \), then \( 1 + 1 + 3 \) represents the solution that gives the first slab to the first processor, the second slab to the second processor and the remaining three slabs to the third processor. Enumerating all possible partitioning to get the optimal one is a feasible solution but can be very computationally expensive for large \( n \) and \( P \). We use a slightly different algorithm for the computations that follows, we choose the permutation with the smallest square difference from the average.

In order to show how our approach works in practice, let us work out the example of using our load balancing example to divide the neghip dataset (the negative potential of a high-potential iron protein of 66\(^3\) resolution) for four processors. Here we assume the number of superslices to be 16, and the number of supervoxels to be 64 (equivalent to a level 4 PARC decomposition). Using a voxel threshold of 10-200 (out of a range up to 255), we get the following 16 supervoxel count for each slab, out of the 1570 total full supervoxels:

\[
12, 28, 61, 138, 149, 154, 139, 104, 106, 139, 156, 151, 129, 62, 29, 13
\]

A naive approach load balancing scheme (such as the ones used in other parallel volume renderers) would assign regions of equal volume to each processor resulting in the following partition:

\[
12 + 28 + 61 + 138 = 239 \\
149 + 154 + 139 + 104 = 546 \\
106 + 139 + 156 + 151 = 552 \\
129 + 62 + 29 + 13 = 233
\]

Here processors 2 and 3 have twice as much "work" as processors 1 and 4. Using our metric, we get:

\[
12 + 28 + 61 + 138 + 149 = 388 \\
154 + 139 + 104 = 397 \\
106 + 139 + 156 = 401 \\
151 + 129 + 62 + 29 + 13 = 384
\]

One can see that some configurations will yield better load balancing than others but this is a limitation of the particular space subdivision one chooses to implement, the more complex the subdivision one allows, the better load balancing but the harder it is to implement a suitable load balancing scheme and the associated ray caster. Figure 12 plots the examples just described for the naive approach. Figure 13 shows how well our load balancing scheme works for a broader set of processor arrangements.

These shortcomings of slabs let us to an alternative space decomposition structure previously used by Ma et al. 67, 68,
Figure 12: The graph shows the number of cubes per processor under naive load balancing.

Figure 13: Load balancing measures for our algorithm. The graph shows the number of cubes the processor receives in our algorithm.

Figure 14: Naive load balancing on the Paragon. The graph shows the actual rendering times for 4 processors using the naive load balancing.

Figure 15: Our load balancing on the Paragon. The graph shows the actual rendering times for 4 processors using our load balancing.

the Binary Space Partition (BSP) tree, originally introduced by Fuchs et al. 41.

Figure 16: An example of the partition scheme we used for load balancing. The bottom represents a possible decomposition for 8 nodes. Notice that a cut can be made several times over the same axis to optimize the shape of the decomposition.

6.2.3. The Parallel Ray Casting Rendering Pipeline

Compositing Cluster

The compositing nodes are responsible for regrouping all the sub-rays back together in a consistent manner, in order to keep image correctness. This is only possible because composition is an associative operation, so if we have to sub-ray samples where one ends where the other starts, it is possible to combine their samples into one sub-ray recursively until we have a value that constitutes the full ray contribution to a pixel. [http]

Ma et al. 68 use a different approach to compositing, where instead of having separate compositing nodes, the rendering nodes switch between rendering and compositing. Our method is more efficient because we can use the special structure of the sub-ray composition to yield a high performance pipeline, where multiple nodes are used to implement
the complete pipeline (see Figure 19). Also, the structure of compositing requires synchronized operation (e.g., there is an explicit structure to the compositing, that needs to be guaranteed for correctness purposes), and light weight computation, making it much less attractive for parallelization over a large number of processors, specially on machines with slow communication compared to CPU speeds (almost all current machines).

It is easy to see that compositing has a very different structure than rendering. Here, nodes need to synchronize at every step of the computation, making the depth of the compositing tree a hard limit on the speed of the rendering. That is, if one uses $2^n$ nodes for compositing, and it takes $t_c$ time to composite two images, even without any synchronization or communication factor in, it takes at least $nt_c$ time to get a completely composited image.

Fortunately, most of this latency can be hidden by pipelining the computation. Here, instead of sending one image at a time, we can send images continuously into the compositing cluster, and as long as we send images at a rate lower than one for every $t_c$ worth of time, the compositing cluster is able to composite those at full speed, and after $mt_c$ times, the latency is fully hidden from the computation. As can be seen for our discussion, this latency hiding process is very sensitive to the rate of images coming in the pipeline. One needs to try to avoid “stalls” as much as possible. Also, one can not pipe more than the overall capacity of the pipeline.

Several implications for real-time rendering can be extracted from this simple model. Even though the latency is hidden from the computation, it is not hidden from the user, at least not totally. The main reason is the overall time that an image takes to be computed. Without network overhead, if an image takes $t_r$ time to be rendered by the rendering cluster, the first image of a sequence takes (at least) time $tr + mt_c$ to be received by the user. Given that people can notice even very small latencies, our latency budget for real-time volume rendering is extremely low and will definitely have to wait for the next generation of machines to be build. We present a detailed account of the timings later in this chapter.

Going back to the previous discussion, we see that as long as $t_r$ is larger than $t_c$ we don’t have anything to worry about with respect to creating a bottleneck in the compositing end. As it turns out, $t_r$ is much larger than $t_c$, even for relatively...
small datasets. With this in mind, an interesting question is how to allocate the compositing nodes, with respect to size and topology.

The topology is actually fixed by the corresponding BSP-tree, that is, if the first level of the tree has \( n = 2^h \) images (if one image per rendering node, then \( n \) would be the number of rendering nodes), than potentially the number of compositing nodes required might be as high as \( 2^h - 1 \). There are several reasons not to use that many compositing nodes. First, it is a waste of processors. Second, the first-image latency grows with the number of processors in the compositing tree. Fortunately, we can lower the number of nodes required in the compositing tree by a process known as virtualization. A general solution to this problem is proposed in Section 6.5.

### Types of Parallelism

Due to the fact that each rendering node gets a portion of the dataset, this type of parallelism is called “object-space parallelism”. The structure of our rendering pipeline makes it possible to exploit other types of parallelism. For instance, by using more than a single rendering cluster to compute an image, we are making use of “image-space parallelism” (in PVR, it is possible to specify that each cluster compute disjoint scanlines of the same image; see 114 for the issues related to image-space parallelism). The clustering approach coupled with the inherent pipeline parallelism available in the compositing process (because of its recursive structure) gives rise to a third parallelism type, namely “time-space parallelism” or “temporal parallelism”. In the latter, we can exploit multiple clusters by concurrently calculating sub-rays for several images at once, that can be sent down the compositing pipeline concurrently. Here, it is important for the correctness of the images, that each composition step be done in lockstep, in order to avoid mixing of images. It should be clear by now that there are several advantages to our separation of nodes into our two types.

### 6.3. Lazy Sweep Ray Casting Algorithm

Lazy Sweep Ray Casting is a fast algorithm for rendering general irregular grids. It is based on the sweep-plane paradigm, and it is able to accelerate ray casting for rendering irregular grids, including disconnected and nonconvex (even with holes) unstructured irregular grids with a rendering cost that decreases as the “disconnectedness” decreases. The algorithm is carefully tailored to exploit spatial coherence even if the image resolution differs substantially from the object space resolution.

Lazy Sweep Ray Casting has several desirable properties, including its generality, (depth-sorting) accuracy, low memory consumption, speed, simplicity of implementation and portability (e.g., no hardware dependencies).

The design of our LSRC method for rendering irregular grids is based on two main goals: (1) the depth ordering of the cells should be correct along the rays corresponding to every pixel; and (2) the algorithm should be as efficient as possible, taking advantage of structure and coherence in the data. With the first goal in mind, we chose to develop a new ray casting algorithm, in order to be able to handle cycles among cells (a case causing difficulties for projection methods). To address the second goal, we use a sweep approach, as did Giertsen 45, in order to exploit both inter-scanline and inter-ray coherence. Our algorithm has the following advantages over Giertsen’s:

1. It avoids the explicit transformation and sorting phase, thereby avoiding the storage of an extra copy of the vertices;
2. It makes no requirements or assumptions about the level of connectivity or convexity among cells of the mesh; however, it does take advantage of structure in the mesh, running faster in cases that involve meshes having convex cells and convex components;
3. It avoids the use of a hash buffer plane, thereby allowing accurate rendering even for meshes whose cells greatly vary in size;
4. It is able to handle parallel and perspective projection within the same framework (e.g., no explicit transformations).

#### 6.3.1. Performing the Sweep

Our sweep method, like Giertsen’s, sweeps space with a sweep-plane that is orthogonal to the viewing plane (the \( x-y \) plane), and parallel to the scanlines (i.e., parallel to the \( x-z \) plane). See Figure 20.

Events occur when the sweep-plane hits vertices of the mesh \( S \). But, rather than sorting all of the vertices of \( S \) in advance, and placing them into an auxiliary data structure (thereby at least doubling the storage requirements), we maintain an event queue (priority queue) of an appropriate subset of the mesh vertices.

![Figure 20: A sweep-plane (perpendicular to the y-axis) used in sweeping 3-space.](image-url)
A vertex \( v \) is **locally extremal** (or simply **extremal**, for short) if all of the edges incident to it lie in the (closed) half-space above or below it (in \( y \)-coordinate). A simple (linear-time) pass through the data readily identifies the extremal vertices.

We initialize the event queue with the extremal vertices, prioritized according to the magnitude of their inner product (dot product) with the vector representing the \( y \)-axis ("up") in the viewing coordinate system (i.e., according to their \( y \)-coordinates). We do not explicitly transform coordinates. Furthermore, at any given instant, the event queue only stores the set of extremal vertices not yet swept over, plus the vertices that are the upper endpoints of the edges currently intersected by the sweep-plane. In practice, the event queue is relatively small, usually accounting for a very small percentage of the total data size. As the sweep takes place, new vertices (non-extremal ones) will be inserted into and deleted from the event queue each time the sweep-plane hits a vertex of \( S \).

The sweep algorithm proceeds in the usual way, processing events as they occur, as determined by the event queue and by the scanlines. We pop the event queue, obtaining the next vertex, \( v \), to be hit, and we check whether or not the sweep-plane encounters \( v \) before it reaches the \( y \)-coordinate of the next scanline. If it does hit \( v \) first, we perform the appropriate insertions/deletions on the event queue; these are easily determined by checking the signs of the dot products of edge vectors out of \( v \) with the vector representing the \( y \)-axis. Otherwise, the sweep-plane has encountered a scanline. And at this point, we stop the sweep and drop into a two-dimensional ray casting procedure (also based on a sweep), as described below. The algorithm terminates once the last scanline is encountered.

We remark here that, instead of doing a sort (in \( y \)) of all vertices of \( S \) at once, the algorithm is able to take advantage of the partial order information that is encoded in the mesh data structure. (In particular, if each edge is oriented in the +\( y \) direction, the resulting directed graph is acyclic, defining a partial ordering of the vertices.) Further, by doing the sorting "on the fly", using the event queue, our algorithm can be run in a "lock step" mode that avoids having to sort and re-insert the vertices. (such cases arise, for example, in some applications of scientific visualization on highly disparate datasets.)

### 6.3.2. Processing a Scanline

When the sweep-plane encounters a scanline, the current sweep status data structure gives us a "slice" through the mesh in which we must solve a two-dimensional ray casting problem. (See Figure 21.) Let \( S \) denote the polygonal (planar) subdivision at the current scanline (i.e., \( S \) is the subdivision obtained by intersecting the sweep-plane with the mesh \( S \).) In time linear in the size of \( S \), we can recover the subdivision \( S \) (both its geometry and its topology), just by stepping through the sweep status structure, and utilizing the local topology of the cells in the slice. In our implementation, \( S \) is actually not constructed explicitly, but only given implicitly by the sweep status data structure, and then **locally** reconstructed as needed during the two-dimensional sweep (described below).

The two-dimensional problem is also solved using a sweep algorithm — now we sweep the plane with a sweep-line parallel to the \( z \) axis. Events now correspond to vertices of the planar subdivision \( S \). At the time that we construct \( S \), we could identify those vertices in the slice that are locally extremal in \( S \) (i.e., those vertices that have edges only leftward or rightward incident on them.), and insert them in the initial event queue. (The actual implementation just sorts along the \( x \)-axis, since the extra memory overhead is negligible in 2D.) The **sweep-line status** is an ordered list of the edges of \( S \) crossed by the sweep-line. The sweep-line status is initially empty. Then, as we pass the sweep-line over \( S \), we update the sweep-line status and the event queue at each event when the sweep-line hits an extremal vertex, making insertions and deletions in the standard way. This is analogous to the Bentley-Ottmann sweep that is used for computing line segment intersections in the plane \(^{10}\). We also stop the sweep at each of the \( x \)-coordinates that correspond to the rays that we are casting (i.e., at the pixel coordinates along the current scanline), and output to the rendering model the sorted ordering (depth ordering) given by the current sweep-line status. We have noticed that the choice of data structure used to maintain the sweep-line status can have a dramatic impact on the performance of the algorithm.

See Silva and Mitchell \(^{113}\) for details.
6.4. Parallel Rendering of Irregular Grids

Here, we present a distributed-memory MIMD machine parallelization of the LSRC method. Our parallelization is a distributed-memory parallelization, and each rendering node gets a only portion of the dataset, not the complete dataset.

The need for the parallelization of rendering algorithms for irregular-grid rendering is obvious, given the fact that irregular grids are extremely large (as compared to regular grids), and their rendering is much less efficient. The largest irregular grids currently being rendered are just breaking the 1,000,000 cell barrier, what would be equivalent to a 100-by-100-by-100 regular grid, if only data sample points are taken into account. On the other hand, such a grid requires more than 50MB of memory, when its regular counterpart only needs 1MB. Actually, regular grids of this size can be rendered by inexpensive workstations in real-time (i.e., using the Shear-Warp technique), while the irregular grids of this size would be almost out of reach just a year ago.

Actually, the sizes of irregular grids of interest of computational scientists are larger than one million cells, possibly starting at two times that range. (This is subjective data, obtained by talking to researchers at Sandia National Labs during the summer of 1996). Given that it takes us about 150 seconds to render a 500,000 cell complex, and assuming linear behavior (what is not completely correct) it would take us over 10 minutes to generate images of a 2,000,000 cell complex. What is not an unreasonable amount of time, given that Ma 70 needed over 40 minutes to render a dataset over 10 times smaller.

But our goal is to develop a method that is both faster and scalable to larger and larger dataset. The main reason for this trust is not really current dataset, but those upcoming ones, specially from the new breed of supercomputers, such as the ACSI TeraFlop machine installed at Sandia National Labs. The ACSI machine has orders of magnitude more memory than the current Intel Paragon installed there, even more usable memory (i.e., not taking OS and network overhead into account). This will enable the generation of extremely large grids, possibly in ranges of 10,000,000-100,000,000 cells or larger.

Part of this increase in dataset sizes can be offset by better algorithms, specially by further improvements in our rendering code by complete implementation of our optimization ideas. But our experience with irregular grids, seems to show that only more computing power can really offset the increase in dataset size.

The other main reason for the use of parallel machines comes from the pure size of the datasets. The largest workstations available to us have 1GB-3GB of memory, what is very short of the 300GB-512GB of memory in the ASCI Tflp machine. Several reasons indicate the visualization should be performed locally; the fact that very few workstations with more than a few gigabytes of memory are available; moving 300GB of data in and out at ethernet, or even ATM OC-3 speeds is clearly infeasible; disk transfer rates, even for reasonably large (and expensive) disk arrays are just too slow for this kind of data.

As all of the reasons pointed above for the use of the parallel machines that generated the dataset is not enough, we also need to note that these simulations do not generate a single static volume, but in general, time dependent data is being generated and the time steps can not, in general, be efficiently accessed (for obvious reasons).

With all of this in mind, we present our algorithm for rendering irregular grid data, in place, on distributed-memory MIMD machines.

6.4.1. Previous Work

There has been very little work on rendering irregular grid data on distributed memory architectures. Overall parallel work on rendering irregular grids has received relatively little attention. This might be due to the fact that rendering irregular grids is so much harder than regular grids, that few people ever get to the point of being able to research parallel methods for irregular grids.

Uselton has parallelize his original ray tracing work (presented in 209) in a shared memory multiprocessor SGI, and reported that the implementation scales linearly up to 8 processors. Challinger 13 reports on a parallel algorithm for irregular grids, implemented on a shared-memory BBN-2000 Butterfly. Giertsen 44 has also parallelized his sweep algorithm on a collection of IBM RS/6000, using a master/slave scheme and total data replication in the nodes.

The most interesting work, by our perspective, is Ma 70, where a parallelization technique very similar to the one presented here is proposed. It is unfortunate that he used a sequential ray casting technique that is shown to be at least two orders of magnitude slower than the one we use. Because of this, he did not find any interesting bottlenecks of the parallelization technique.

His technique works by breaking up the original grid into multiple, disjoint cell complexes using Chaco 48, a graph-based decomposition tool developed at Sandia National Labs. Chaco-based decompositions have several interesting and important properties for parallelization of computational methods. It is unclear, the extra overhead of using Chaco has actually any influence on the rendering speed of the parallelization. Here, as in our parallel regular grid method presented in Section 6.2, we divide the nodes into two classes: rendering and compositing nodes. The rendering nodes, compute each ray of an image, creating a set of stencils (the rays may not be completely connected). After each ray is computed, they are sent to the compositing nodes for further sorting and the final accumulation. Each compositing node is assigned a set of rays to be composited. He
reports that because the rendering takes so long, the compos-  
posing phase is negligible and he has not work any further on optimizing it.

6.4.2. Parallel LSRC

Overall the our algorithm is very similar to Ma’s. Continuing in the tradition of our regular grid work and the framework of our PVR system, we divide the nodes into two relevant groups, rendering and compositing nodes. Our differences between our work and Ma’s are actually in the details of the rendering and compositing.

Dataset Decomposition

In order to subdivide the dataset among the nodes, we use a hierarchical decomposition method, with a similar flavor to our load balancing scheme for regular grids. Starting with the bounding box of the complete cell complex, we start making cuts in this box, taking two things into account: the aspect ratio of the cuts, and the number of vertices. At every step, we cut along the largest axis in such a way as to break the number of vertices in half, in each stage of the cutting. Because cells might belong to more single of these convex space decomposition “boxes”, we assign the cell to the box that has most of it (e.g., in the number of vertices, with ties broken in some arbitrary, but consistent way).

The obvious now, is just to assign each processor to a each box. This is a way to minimize the total rendering time of the complete irregular grid. Unfortunately, it is not clear that this is the right thing, given that one might want to create a rough picture of the grid fast, then wait for more complex rendering. In the future we expect to be able to create a scattered decomposition, that will have better properties in creating approximate renderings of the grids.

With the decomposition method just proposed, each processor should have roughly the same number of primitives, each of which, approximately confined to a rectangular grid of almost bounded aspect ratio (because of the largest-axis cutting).

Rendering

The rendering performed at each node is just a variation of our sequential technique presented in Section 6.3. This is just a single significant difference, instead of generating an image, every node generates a stencil data-structure. Of course, all nodes work concurrenctly on generating stencil scan-lines.

The stencil representation of a scan-line is just a linked-list of color and depth of cells, who have been lazily composited. That is, if two stencils shared an end point (e.g., \((a, b)\) and \((b, c)\), they are composited into a single stencil \((a, c)\), representing the whole region. In the end of a scan-line rendering computation, each node potentially has a collection of stencils. Because of the process of decomposing the dataset among the nodes, it is expected the stencil fragmentation is low. This is necessary in order to enable fast communication for compositing.

Compositing

One solution for compositing would just to copy Ma’s technique, where nodes are responsible for certain scan-lines. This way the rendering nodes could just send its collection of stencils for further sorting in the compositing nodes. In our case, we try to achieve better performance by creating a tree of compositing nodes (such as the one we use for the regular case). Every compositing node is responsible for a certain region of space (i.e., one of the original box decompositions proposed above), that belongs to a global BSP-tree.

It is the responsability of the rendering nodes to respect the BSP-tree boundaries and send the data to the correct compositing nodes, possibly breaking stencils that are span across boundaries.

Once the data of each scan-line is received in the compositing nodes, the final depth sorting can be efficiently performed by merging the stencils into a complete image. An efficient pipeline scheme can be implemented on a scan-line by scan-line basis, with similar good properties as the one implemented image-by-image for the regular grid case.

6.5. General BSP-tree Compositing

A simple way of parallelizing rendering algorithms is to do it at the object-space level: i.e., divide the task of rendering different objects among different rendering processors, and then compose the full images together. A large class of rendering algorithms (although not all), in particular scan-line algorithms, can be parallelized using this strategy. Such parallel rendering architectures, where renderers operate independently until the visibility stage, are called sort-last (SL) architectures. A fundamental advantage of SL architecture is the overall simplicity, since it is possible to parallelize a large class of existing rendering algorithms without major modifications. Also, such architectures are less prone to load imbalance, and can be made linearly scalable by using more renderers. One shortcoming of SL architectures is that very high bandwidth might be necessary, since a large number of pixels have to be communicated between the rendering and compositing processors. Despite the potential high bandwidth requirements, sort-last has been one of the most used, and successful parallelization strategies for both volume rendering and polygon rendering, as shown by the several works published in the area.

Here we present a general purpose, optimal compositing machinery that can be used as a black box for efficiently parallelizing a large class of sort-last rendering algorithms. We consider sort-last rendering pipelines that are based on separating the rendering processors from the compositing processors, similar to what was proposed previously by Molnar.
The techniques described in this paper optimize overall performance and scalability without sacrificing generality or the ease of adaptability to different renderers. Following Molnar, we propose to use a scan-line approach to image composition, and to execute the operations in a pipeline as to achieve the highest possible frame rate. In fact, our framework inherits most of the salient advantages of Molnar’s technique. The two fundamental differences between our pipeline and Molnar’s are:

1. instead a fixed network of Z-buffer compositors, our approach uses a user-programmable BSP-tree based composition tree;
2. we use general purpose processors and networks, instead of Molnar’s special purpose Z-comparators arranged in a tree.

In our approach, hidden-surface elimination is not performed by Z-buffer alone, but instead by executing a BSP-tree algorithm, shown in Figure 23, please see Ramakrishnan and Silva 7.

/* The algorithm marks the beginning of partitions in the subtree of \( G \) rooted at \( u \). If more vertices, can be added to the root partition, the algorithm returns the size of the root partition.
Otherwise, the algorithm returns 0. */
1. if \( \text{arity}(u) = 2 \) then /* \( u \) is a binary vertex */
2. \( w_1 := \text{partition(left\_child}(u)) \);
3. \( w_2 := \text{partition(right\_child}(u));\)
4. \( w := w_1 + w_2 + 1;\)
5. if \( (w > K) \) then
6. if \( (w_1 \leq w_2) \) then
7. Mark right\_child(u) as start of new partition
8. \( w := w_1 + 1;\)
9. else
10. Mark left\_child(u) as start of new partition
11. \( w := w_2 + 1;\)
12. else if \( \text{arity}(u) = 1 \) then /* \( u \) is a unary vertex */
13. \( w := \text{partition(child}(u));\)
14. else /* \( u \) is a leaf */
15. \( w := 1;\)
16. if \( (w = K) \) then
17. Mark \( u \) as a start of new partition
18. return(0);
19. else
20. return(w);

Figure 23: Algorithm partition

6.5.1. Optimal Partitioning of the Compositing Tree

We can view the BSP tree as an expression tree, with compositing being the only operation. In our model of compositing clusters, evaluation of the compositing expression is mapped on to a tree of compositing processes such that each process evaluates exactly one sub-expression. See Figure 22 for an illustration of such a mapping. The actual ordering of compositing under a BSP-tree depends not only on the position of the nodes, but also on the viewing direction. So, during the execution phase, a specific ordering has to be obeyed. Fortunately, given any partition of the tree, each subtree can still be executed independently. Intuitively, correctness is achieved by having the nodes “fire up” in a on-demand fashion.

Such a decompositing is based on a model of the cost of the subtrees. For details on this, and the partitioning algorithm, shown in Figure 23, please see Ramakrishnan and Silva 97.

Practical Considerations

Algorithm partition provides a simple way of given a BSP-tree, and a performance requirement, given in terms of the frame rate, how to divide up the tree in such a way as to optimize the use of processors. Several issues, including ma-
machine architecture bottlenecks, such as synchronization, interconnection bandwidth, mapping the actual execution to a specific architecture (e.g., a mesh-connected MIMD machine) were left out of the previous discussion. We now describe how Algorithm partition can be readily adapted to account for some of the above issues in practice.

**Compositing Granularity:** Note that there is nothing in the model that requires that full images be composited and transferred at a time. Actually, one should take into consideration when determining the unit size of work, and communication, hardware constraints such as memory limitations, and bandwidth requirements. So, for instance, instead of messages being a full image, it might be better to send a pre-defined number of scan-lines. Notice that in order for images of arbitrary large size to be able to be computed, the rendering algorithm must also be able to generate the images in scan-line order.

**Communication Bandwidth:** Of course, in order to achieve the desired frame rate, enough bandwidth for distributing the images during composition is strictly necessary. Given \( p \) processors, each performing \( k \) compositing operations, the overall aggregate bandwidth required is proportional to \( p(k+2) \). It should be clear that as \( k_{\text{max}} \) increases, the actual bandwidth requirement actually decreases (both for the case of a SL-full, as well as a SL-sparse architecture) since as \( k_{\text{max}} \) increases the number of processors required decreases. This decrease in bandwidth is due to the fact that compositing computation are performed locally, inside each composite processor, instead of being sent over the network. If one processor performs exactly \( k_{\text{max}} \) compositing operations, it needs \( k_{\text{max}} + 2 \) units of bandwidth, as opposed to \( 3k_{\text{max}} \) when using one processor per compositing operation— a bandwidth savings of almost a factor of three!

Another interesting consideration related to bandwidth is the fact that our messages tend to be large, implying that our method operates on the best range of the message size versus communication bandwidth curve. For instance, for messages smaller than 100 bytes the Intel Paragon running SUNMOS achieve less than 1 MB/sec bandwidth, while for large messages (i.e., 1MB or larger), it is able to achieve over 160MB/sec. (This is very close to 175MB/sec, which is the peak hardware network performance of the machine.) As will be seen in Section 6.5.2, our tree execution method is able to completely hide the communication latency, while still using large messages for its communication.

**Latency and Subtree Topology:** As will be seen in Section 6.5.2, the whole process is pipelined, with a request-based paradigm. This greatly reduces the overhead of any possible synchronization. Actually, given enough compositing processors, the overall time is only dependant on the performance of the rendering processors. Also, note that the actual shape of the subtree that a given processor gets is irrelevant, since the execution of the tree is completely pipelined.

**Architectural Topology Mapping:** We do not provide any mechanism for optimizing the mapping from our tree topology to the actual processors in a given architecture. With recent advancements in network technology, it is much less likely that the use of particular communication patterns improve the performance of parallel algorithms substantially. In new architectures, the point-to-point bandwidth in access of 100–400 MB/sec are not uncommon, while in the old days of the Intel Delta, it was merely on the order of 20 MB/sec. Also, network switches, with complex routing schemes, are less likely to make neighbor communication necessary. (Actually, the current trend is not to try to exploit such patterns since new fault-handling and adaptive routers usually make such tricks useless.)

**Limitations of Analytical Cost Model:** Even though we can support both SL-full and SL-sparse architecture, our...
model does not make any distinction of the work that a given compositing processor is performing based on the depth of its compositing nodes. This is one of the limitations of our analytical formulation. However, the experimental results indicate that this limitation does not seem have any impact on the use of our partitioning technique in practice. Actually, frame-to-frame differences might diminish the concrete advantage of techniques that try to optimize for this fact.

6.5.2. Optimal Evaluation

In the previous section, we described techniques to partition the set of compositing operations and allocate one processor to each partition, such that the various costs of the compositing pipeline can be minimized. We now describe efficient techniques for performing the compositing operations within each processor.

Space-Optimal Sequential Evaluation of Compositing Trees

Storage is the most critical resource for evaluating a compositing tree. We need 4MB of memory to store an image of size $512 \times 512$, assuming 4-bytes each for RGB and $\alpha$ values per pixel. Naive evaluation of a compositing tree with $N$ nodes may require intermediate storage for up to $N$ images.

We now describe techniques, adapted from register allocation techniques used in programming language compilation, to minimize the total intermediate storage required. Figure 24a shows a compositing tree for compositing images $I_1$ through $I_6$. We can consider the tree as representing the expression

$$\left( I_1 \oplus \left( I_2 \oplus \left( I_3 \oplus I_4 \right) \right) \right) \oplus \left( I_5 \oplus I_6 \right)$$

(11)

where $\oplus$ is the compositing operator. Since images $I_1$ through $I_6$ are obtained from remote processors, we need to copy these images locally into intermediate buffers before applying the compositing operator. The problem now is to sequence these operations and reuse intermediate buffers such that the total number of buffers needed for evaluating the tree is minimized.

We encounter a very similar problem in a compiler, while generating code for expressions. Consider a machine instruction (such as integer addition) that operates only on pairs of registers. Before this operation can be performed on operands stored in the main memory, the operands must be loaded into registers. We now describe how techniques to generate optimal code for expressions can be adapted to minimize intermediate storage requirements of a compositing process. The number of registers needed to evaluate an expression tree can be minimized, using a simple tree traversal algorithm pages 561–562. Using this algorithm, the compositing tree in Figure 24a can be evaluated using 3 buffers. In general, $O(\log N)$ buffers are needed to evaluate a compositing tree of size $N$. However, by exploiting the algebraic properties of the operations, we can further reduce the number of buffers needed—to $O(1)$. Since $\oplus$ is associative, evaluating expression (11) is equivalent to evaluating the expression:

$$\left( (I_1 \oplus I_2) \oplus \left( I_3 \oplus I_4 \right) \right) \oplus \left( I_5 \oplus I_6 \right)$$

(12)

The above expression is represented by the compositing tree in Figure 24b, called an associative tree. The associative tree can be evaluated using only 2 buffers.

Again, for full details, we refer the reader to the full paper.

6.5.3. Implementation

In this section, we sketch the implementation of our compositing pipeline. We implemented our compositing backend in the PVR system. PVR is a high-performance volume rendering system, and it is freely available for research purposes. Our main reason for choosing PVR was that it already supported the notion of separate rendering and compositing clusters, as explained in Chapter 3. The basic operation is very simple. Initially, before image computation begins, all compositing nodes receive a BSP-tree defining the compositing operations based on the object space partitioning chosen by the user. Each compositing node, in parallel, computes its portion of the compositing tree, and generates a view-independent data structure for its part. Image calculation starts when all nodes receive a sequence of viewpoints.

The rendering nodes, simply run the following simple loop:

```c
For each (viewpoint v)
    ComputeImage(v);
    p = WaitForToken();
    SendImage(p);
```

Notice that the rendering nodes do not need any explicit knowledge of parallelism; in fact, each node does not even need to know, a priori, where its computed image is to be sent. Basically, the object space partitioning and the BST-tree takes care of all the details of parallelization.

The operation of the compositing nodes is a bit more complicated. First, (for each view) each compositing processor computes (in parallel, using its portion of the compositing tree) an array with indices of the compositing operations assigned to it as a sequence of processor numbers from which it needs to fetch and compose images. The actual execution is basically an implementation of the perfect scheduling scheme proposed here, with each read_request being turned into a PVR_MSG_TOKEN message, where the value of the token carries its processor id. So, the basic operation of the compositing node is:

```c
For each (viewpoint v)
    CompositeImages(v);
    p = WaitForToken();
    SendImage(p);
```
Notice that there is no explicit synchronization point in the algorithm. All the communication happens bottom-up, with requests being sent as early as possible (in PVR, tokens are sent asynchronously, and in most cases, the rendering nodes do not wait for the tokens), and speed is determined by the slowest processor in the overall execution, effectively pipelining the computation. Also, one can use as many (or as few) nodes one wants for the compositing tree. That is, the user can determine the rendering performance for a given configuration, and based on the time to composite two images it is straightforward simple to scale our compositing back-end for his particular application.
PART THREE
Case Studies

7. The PVR System

This manuscript is a pre-publication version of “PVR: High Performance Volume Rendering”, C. Silva, A. Kaufman, and C. Pavlakos.

Cláudio T. Silva* Arie E. Kaufman* Constantine Pavlakos‡
*State University of New York at Stony Brook ‡Sandia National Laboratories

7.1. Introduction

Volume rendering is a powerful computer graphics technique for the visualization of large quantities of 3D data. It is specially well suited for the visualization of three dimensional scalar and vector fields. Fundamentally, it works by modelling the volume as cloudy-like cells composed of semi-transparent material which emits its own light, partially transmits light from other cells and absorbs some incoming light. (See sidebar Volume Rendering for details).

In order to allow researchers and engineers make effective use of volume rendering to study complex physical and abstract structures, a coherent, powerful, easy to use visualization tool is needed. Furthermore, such a tool should allow for interactively visualization, ideally with support for user-defined “computational steering”.

There are several issues and challenges in developing such visualization tools. (1) So much as the latest volume rendering acceleration techniques running on top-of-the-line workstations, it still takes a few seconds to a few minutes to volume render images. This is clearly far from interactive. With the advent of larger parallel machines, better scanners and instrumentation, larger and larger datasets (typically from 32MB to 512MB, but with sizes as high as 16GB) are being generated, some of which would not even fit in memory of a workstation class machine. (2) Even if rendering time is not a major concern, big datasets may be expensive to hold in storage, and extremely slow to transfer to typical workstations over network links.

These issues lead to the question of whether the visualization should be performed directly on the parallel machines which is used to generate the simulation data or sent over to a high performance graphics workstation for post-processing. First, if the visualization software was integrated in the simulation software, there would be no need for extra storage and visualization could be an active part of the simulation. Second, large parallel machines can render these datasets faster than workstations can, possibly in real-time or at least achieving interactive frame-rates. Finally, the integration of simulation and visualization in one tool (when possible) is highly desirable, because it allows users to interactively “steer” the simulation, and possibly terminate (or modify parameters in the) simulations instead of performing painfully long simulations on extremely expensive machines, with high cost of storage and transmission, only to find out at post-processing that the simulations are wrong or uninteresting.

In this paper we introduce the PVR system, currently being developed under a collaboration between Sandia National Laboratories and the State University of New York at Stony Brook. PVR is a component approach to building a distributed volume visualization system. On its topmost level it provides a flexible and high performance client/server volume rendering architecture to the user with a unique load balancing scheme which provides a continuum of cost/performance parameters that can be used to optimize rendering speed. The original goals of PVR were to achieve a level of portability and performance for rendering beyond that of other available systems and to provide a platform that can be used for further development.

But PVR is more than a rendering system, its components were specially designed to be user-extensible, in order to allow for user defined computational steering (that is, the user can easily add his own computational code to PVR and just link in our rendering library). Using PVR, it is much simpler to build portable, high performance, complex distributed visualization systems (Figure 25).

The rest of the paper introduces the PVR client/server architecture and its components, with emphasis on its application to volume rendering. Details on how to achieve computational steering with PVR are scattered across this paper.


7.2. The PVR System

It is well known that system complexity always limits the reliability of large software projects. Distributed systems exacerbate this problem with the introduction of asynchronous and non-local communication. With all of this in mind, we use a component approach to developing PVR. PVR attempts to provide just enough functionality in the basic system to allow for the development of large and complex computational steering and visualization applications. It is based on a client/server architecture, where there are coupled rendering/computing servers on one side, and on the other the user acts as a client (from his workstation).

The PVR client/server architecture is implemented in two main components: the pvrs, which runs in the user workstation, and the PVR renderer, running in parallel machines. The renderer is implemented as a library and it allows for easy integration of user defined code that can share the same processors as the rendering. Communication across applications with PVR, are performed using the PVR protocol, and in our implementation communication is handled by separate UNIX processes (see Figure 26).

7.2.1. The pvrs

The pvrs is an augmented Tcl/Tk shell. It provides a single new object to the user, the PVR session. We chose to use Tcl/Tk as the system glue. Tcl (Tool command language) is a script language designed to be used as a generic language in application programs. It is easily extendable with new user commands (in C or Tcl) and coupled with the graphical environment Tk, it is a powerful graphical user interface system. The use of the Tcl/Tk, which are well-designed, debugged application language and graphical environment contribute to reducing the overall system complexity.

The PVR session is an object (such as the Tk objects). It contains attributes and corresponding methods (used to change those attributes). One of the most important attributes is the one that binds a session to a particular parallel machine. Figure 26 contains three sessions, two on acomo.cs.sandia.gov (a large Intel Paragon XP/S with over 1840 nodes running SUNMOS installed at Sandia) and one on parxp2.ams.sunysb.edu (a small Intel Paragon with 110 nodes running Intel’s version of OSF/I installed at Stony Brook). The system is designed to handle multiple sessions using the same protocol with machines running different operating systems.

As part of its attributes a session specifies the number of nodes it needs, and the parameters that are passed to those nodes. Several pieces of informations are interactively exchanged between the pvrs and the PVR renderer, such as rendering configuration information, rendering commands, sequences of images, performance and debugging information.

There is a high amount of flexibility in the specification of the rendering. Not only simple rendering elements, such as changing transformation matrices, transfer functions, image sizes, datasets can be specified, but there are commands to specify (in a high level format) the complete parallel rendering pipeline (see sidebar Parallel Volume Rendering for details). With these parameters in hand, the pvrs can be used to specify almost arbitrary scalable rendering configurations (see Section 7.2.4).

The pvrs is implemented as a single process (used to make ports easier) in about 5,000 lines of C code. We have augmented the Tcl/Tk interpreter with TCP/IP connection capabilities (some versions of Tcl/Tk have this build in). Due to the need of several concurrent sessions, all the communication is performed asynchronously. We use the Tk_CreateFileHandler() routine to arbitrate between input from the different sessions (a UNIX select call and polling could be used instead, but would make the code harder to understand and overall more complex). Sessions work as interrupt driven commands, responding to requests one at a time (every session can receive events from two sources at the same time, the user keyboard and the remote machine). Locking and disabling interrupt are needed to ensure consistency inside critical sessions.

The overall structure of the code allows for user augmentation of a session’s functionality either by external or internal means. External augmentation is the one that can be performed without re-compilation (such as the one used by the user interface to show the images as they are received asynchronously from the remote parallel server). Internal augmentation requires changes to the source code. The source code is structured to allow for simple addition of new functionality. Only a single file needs to be changed to add a new session method (if it changes the Resource Database, two files need to be changed). New commands are added using Tcl conventions (see Part 3 of Ousterhout’s Tcl/Tk book).

Every PVR message is sent either as a single fixed length message, or as two messages (the first is used to specify the size of the second). This is used to make redirection easier and to achieve optimal performance under different configurations. Look up tables are set up with actions to be taken on the arrival of each message type. This setup makes additions to the PVR protocol very simple.

7.2.2. The PVR renderer

The PVR renderer is the piece of PVR that runs remotely on the parallel machine (see Figure 26). It is composed of several components, the most complex being the rendering itself. In order to start up multiple the parallel processes at the remote machine, we use pvd, the PVR daemon. This daemon runs in the parallel machine. It waits on a well-known port for connection requests. Once a request for opening a new session is made it forks a handling process that will be responsible for allocating processors and communicating with the session on the client. In the remote machine, the handling
process allocates the computing nodes, and runs the renderer code on them. The connection process is illustrated in Figure 27. One pvrd can allocate several processes, once it is killed, it kills all its children before exiting.

The renderer is the code that actually runs on the parallel nodes. The overall structure of the code resembles an SIMD machine, where there are high-level commands and low-level commands. There is one master node (similar to the microcontroller on the CM-2 machines), and several slave nodes. The functions of the slaves are completely dependent on the master. The master receives commands from the pvrs, translates them, and takes the necessary actions, including changing the state of the slaves and sending them a detailed set of instructions.

For flexibility and performance, the method of sending instructions to the nodes are through action tables (like SIMD microcode). In order to ask the node to perform some action, the master broadcast the address of the function to be executed. Upon receiving that instruction, the slaves execute that particular function. With this method, it is very simple to add new functionality, because any new added functionality can be performed locally, without the need to change global files. Also, every function can be optimized independently, with its own communication protocol. One shortcoming of this communication method (as in SIMD machines) is that one has to be careful with non-uniform execution, specially because the Intel NX communication library (both OSF and SUNMOS have support for NX) has limited functionality for handling nodes as groups (e.g., in setting up barriers with NX it is impossible to select a group from the totality of the allocated nodes).

The master intrinsically divides the nodes into clusters. Each cluster has specialized computational task, and multiple clusters can cooperate in groups to achieve a large task.

All that is necessary for cluster configuration is that the basic functions be specified in user-defined libraries that are linked in a single binary. During runtime, the user can use the master to reconfigure his clusters accordingly to his immediate goal. The pvrs can be used to interactively send such commands. As an example of the use of such scheme, see Figure 28, where the rendering configuration for PVR’s high performance volume renderer is specified.

In order to achieve user-defined computational steering, one can use this clustering paradigm. It will usually be necessary to add one’s functionality to the action tables (e.g., linking the computational code with PVR dispatching code), and also add extra options to the pvrs (usually through the set command) for modifying the relevant parameters interactively.
PVR’s volume rendering code was the inspiration for this overall code organization and is a very good application to demonstrate its features. Because in this paper our focus is on describing PVR, and not on the actual volume rendering code, we only sketch the implementation to give insight on how to add your own code (for possible computational steering) to PVR and to give you enough information for effective use of PVR’s rendering facilities.

7.2.3. Volume Rendering Pipeline

The PVR rendering pipeline is composed of three types of nodes (besides the master, of course). There are the rendering nodes, compositing nodes, and collector nodes (usually just one), (see Figure 28). This specialization is necessary for optimal rendering performance and flexibility. All the clusters work in a simple dataflow mode, where data moves from top to bottom in pipeline fashion. Every cluster has its own fan-in and fan-out number and type of messages. The master configures (and re-configures) the overall dataflow using a set of user-defined and automatic load balancing parameters.

At the top level are the rendering clusters. The nodes in a rendering clusters are responsible for the resampling and shading a given volume dataset. In general the input is a view matrix, and the output is a set of sub-images, each of which is a related to a node in the compositing binary tree. The master can use multiple rendering clusters working on the same image, but disjoint scanlines in order to speed up rendering. Once the subimages are computed, they are passed down in the pipeline to the compositing clusters.

The compositing clusters are organized in a binary tree structure, matching that of the corresponding compositing tree. The number of compositing nodes can actually be different, as we can use virtualization to fake more processors than allocated. Images are pipelined down the tree, with every iteration combining the results of compositing until finally all the pixels are a complete depth-ordered sequence. Those pixels are converted to RGB format and sent to the collector node(s) (at this time, we just use a single collector node).

The collector node receives RGB images from the compositing nodes, compresses them using a simple run-length encoding scheme (very fast compression is necessary). Finally the images are either sent over to the pvrsh for user viewing (or saving), or locally cached on the disk (it can also be specified that images are trashed for performance analysis purposes).

The previous discussion is over simplistic. There are several performance issues, related to CPU speed, synchronization and memory usage that have not been discussed. For more complete details, we refer the interest reader to Cláudio Silva’s Ph.D. dissertation

7.2.4. Rendering with PVR

Figure 29 shows a simple PVR program. Several important features of PVR are demonstrated. In particular, the seamless integration with Tcl/Tk, the flexible load balancing scheme, and the interactive specification of parameters. The set command can have several options (in the Figure 29 they are usually specified in multiple lines, but all can be specified in a single line). For instance, -imagesz specifies the size of the images that are outputed by the system.

The -cluster and -group options are unique to PVR and its flexible load balancing scheme. With both of these options, the relative sizes of the rendering and compositing clusters can be specified together with the image calculation allocation. Several scalability strategies can be used, for instance, a rendering cluster needs to be large enough to hold the entire dataset and at least a copy of the image to be calculated. By increasing the size of the cluster, the amount of memory per node decreases. By grouping clusters (using -group), the number of scanlines a given cluster is responsible for decreases, lowering both the memory and the computational cost, thus speeding up image calculation. The same commands can be used to configure compositing clusters. The scalability parameters for compositing clusters is very different than for rendering clusters, because of the different nature of the task. Compositing nodes need memory to hold two copies of the images, what can be quite large (our current parallel machines only have between 16MB to 32MB RAM, this might not be a problem in the near future), and also compositing has very high synchronization cost that grows as the number of nodes grow. Currently, the only need for multiple compositing clusters is due to the need of more memory for large images (such as 1024-by-1024).

7.2.5. DVEs

DVEs can be easily developed by making use of the client/server metaphor. DVE developed using Tcl/Tk are very portable, as Tcl/Tk has ports for almost all the operating systems available, and TCP/IP (our communication protocol) is virtually universal.

Figure 30 shows a simple prototype GUI developed at Sandia. The complete interface is written in Tcl/Tk. The user is able to specify all the necessary rendering parameters in the right window (including image size, transfer function, etc.) and the load balancing parameters in the left window. This simple interface only uses a single session at this time, but more functionality is being added to the system.

Using the prototype GUI, users are able to add their own functionality to the system as needed. This flexibility not only makes the system more usable, because redundant bells and whistles can be discarded, but also new functionality can be easily added. The use of a portable and well-documented windows interface (e.g., Tk) is imperative. Not only users avoid having to learn yet another programming language.
Figure 28: The host receives high level commands that are translated into virtual microcode by the action tables. For rendering, the high level commands are for the generation of animations by rotations and translations, that are translated into simple transformation matrices commands. The rendering clusters perform rendering in parallel. The collector receives and groups images back together and sends an ordered image sequence to the client application.

and graphical toolkit, but the use of Tk saved us a lot of implementation and documentation cost (Tcl/Tk is widely used and highly documented). Another important feature of Tcl/Tk for the development of prototypes is that it is freely available, enabling us to do the same for PVR.

7.3. Discussion

7.3.1. Related Work

The Shastra project at Purdue has developed tools for distributed and collaborative visualization. Their system implements parallel volume visualization with a mix of image space and object space load balancing, but few details of the scheme are given. They report using up to 4 processors for computation, what makes it hard to evaluate the systems usability in a massively parallel environment. Rowlan et al. describes a distributed volume rendering system implemented on the IBM SP-1. Their system has several of the same characteristics as ours. They also separate rendering and compositing nodes to increase performance and provide a Front-End GUI. Another cousin of our system is DISCOVER, developed at National Cheng-Kung University (Taiwan). Their system was customly developed for medical imaging applications and provides mechanisms for the use of remote processor pools.

7.3.2. Visualization Servers

One use of our parallel renderer is as a visualization server for parallel processes. The basic idea is to pre-allocate a set of nodes that can be time-shared by multiple users for visualizing their data. Because of our novel use of pipelining, this can be achieved fairly efficient and with minimal overhead.

Actually, our system architecture is also suitable for time-varying data. When rendering time-varying data, we add a permanent caching clusters to the pipeline in Figure 28, that is responsible for distributing the volume data to the rendering nodes efficiently. The caching cluster is used to hide I/O latency from disk (or other sources). This way the user can visualize his data for as long as a new version comes along. Handling data that changes too rapidly (e.g., faster than we can move it and render) is not possible as it would require large amounts of buffering.

7.3.3. Performance and Results

The current version of PVR is about 25,000 lines of C and Tcl/Tk code. It has been used at Brookhaven National Labs, Sandia National Labs and Stony Brook for the visualization of large datasets. We have demonstrated the capability of rendering a 500,000,000 bytes dataset (the CT visible human data from the National Institute of Health) in approximately 5 seconds/frame. Actually, PVR was demoed in the Sandia booth during Supercomputing ’95. Our plans are to use render the full RGB visible human (14,000,000,000 bytes) by the end of the year. (Parallel I/O will be a must, currently the 500MB visible human takes 15 minutes of disk I/O).

7.3.4. Further Development

The idea of developing PVR started out of frustration trying to use network of workstations and the Paragon as render-
Figure 29: A set of PVR rendering commands. The commands can be put in a file and executed in batch, or can be typed interactively on the keyboard (or mixed). Tcl/Tk code (such as “stat.tcl”) can be written to take care of portions of the actions.

It was always clear that a pure distributed approach to building rendering environments would be much more powerful than special rendering tools with parallel capabilities. Even though we have completed a usable and efficient system, there is still a long wish list, in both the research and development front.

We are currently working on making the system stable enough for large scale availability. With that in mind we are currently working on creating a complete DVE (using VolVis as a starting point) on top of PVR. One of the challenges is how to integrate resource allocation and admission control in our DVE.

Some functionality is missing from PVR and needs to be incorporated. The most important is probably the support for multiple data sets in a session. This would make the load balancing scheme much more complicated, and simple heuristics might not generate well balanced decomposition schemes. If the volumes are allowed to overlap (as in VolVis), the problem is even harder, and the solution seems to require having special composite nodes that perform the sorting at the end of the pipeline. It might be necessary to have a reconfiguration phase each time a new volume is introduced in the picture. It is not clear yet how this can be done efficiently.

A simpler change is to add support in the PVR renderer for non-homogeneous processors. One just needs to change

```bash
:brain set -dataset brain.slc
:brain set -cluster r,16 -group 0,0,1,1
:brain set -cluster c -group 0,0
:brain set -cluster r,16 -group 0,1,2,3,4
:brain render rotation 1,1,1 0,359:360
```
the load balancing scheme slightly by normalizing the number of PARC sub-cubes per processor with their relative performance parameters. Finally, we hope to look in the future for ways to perform real-time manipulation rendering, where the user can just move a mouse and see the picture changing in real-time with minimal lag. For this, we suspect the work done in eliminating virtual reality lag may help.

7.4. Conclusions

In this article we have introduced the PVR system. Here are some of the key features in our system:

- **Transparency** - PVR hides most of the hardware dependencies from the DVEs and the user.
- **Performance** - PVR provides a high speed pipelined ray caster with a unique load balancing scheme and mechanisms to fine tune performance for any given machine configuration.
- **Scalability** - All the algorithms used in the system were carefully chosen to be gracefully scalable. Not only with respect to the machine size, but special care was taken to allow for grown in dataset size and image size.
- **Extensibility** - The PVR architecture can be easily extended, making it easy for the DVE to add new functionality. Also, it is fairly easy for the user to add new functionality to the PVR shell and its corresponding kernel, allowing for user defined “computational steering” coupled with visualization.

PVR introduces a new level of interactivity to high performance visualization. Larger DVEs can be built on top of PVR, and yet be portable across several architectures. These DVEs that use PVR are given the opportunity to make effective use of available processing power (upto to a few hundred processors), giving a range of cost/performance to end users. This is particularly important in the scientific research community, since most often the question is not how fast, but how much. PVR provides a strong foundation for building cost effective DVEs.

As far as the user interface is concerned, PVR introduces a much simpler way to create it. No longer one has to spend time coding in X/MOTIF (or Windows) to create the desired user interface. The Tcl/Tk combination is much simpler, gives more flexibility, and is closely as powerful as the other alternatives. Tcl/Tk is becoming as popular as UNIX shell programming. Different sites should be able to easily create and modify their own systems.

Acknowledgments

We would like to thank Maurice Fan Lok, who as a M.S. student at Stony Brook made fundamental contributions to PVR by co-writing the first version during 1994/95. Brian Wylie deserves special thanks for continual support of the project and for the development of the user interface. We thank Tzi-cker Chiueh, Pat Crossno, Steve Dawson, Juliana Freire, Ron Peierls, and Amitabh Varshney for useful discussions about the PVR system and for help with this paper. The port of PVR to SUNMOS was only possible due to the help of Kevin McCurley, Rolf Riesen, Lance Shuler from Sandia and Edward J. Barragy from Intel. The MRI data set is courtesy of Siemens, Princeton, NJ. C. Silva is partially supported by CNPq-Brazil under a PhD fellowship and by Sandia National Labs. A. Kaufman is partially supported by the National Science Foundation under grants CCR-9205047 and DCA 9303181 and by the Department of Energy under the PICS grant.

Electronic Information

Current information on PVR (including images, animations, related publications, etc.) is kept in http://cg.ams.sunysb.edu/~pvr and http://www.cs.sandia.gov/VIS. The source code (Paragon version only) is currently available upon request. Please send mail to csilva@ams.sunysb.edu.

Appendix: Parallel Volume Rendering

There are basically three different orthogonal types of parallelism that can be exploited with volume rendering. In the PVR system we exploit all of them:

**Image Space Parallelism.**

In image space parallelism parts of a single image are divided into multiple processors for concurrent image generation. For volume rendering (and in general for computer graphics) this is the simplest form of parallelism, as no subdivision of the volume itself need to be performed and the volume never needs to be updated. This is usually the type of parallelism implemented on shared memory machines. The main shortcoming of this method (for distributed memory machines) is that large datasets can not be rendered using this type of parallelism alone.

**Object Space Parallelism.**

In object space parallelism parts of the volume are divided into multiple processors, each computes a sub-image, that is later regrouped. The main shortcomings are the higher need for communication and synchronization among the processors (parallel machines still have slow communications with respect to processor speed), Ma et al. describes an efficient algorithm for re-grouping the images back together to form a single correct image. In statically partitioning the volume dataset, one has to be careful to give every processor the same amount of work.

**Time Space Parallelism.**

In time space parallelism, multiple images are computed at the same time. This exploits the fact that the rendering process can be easily divided into multiple disjoint phases.
8. Building a Linux-based Parallel Machine

See tutorial website.

Acknowledgments

Numerous individuals have contributed to all parts of the material presented here.

We want to thank the members of the Personal Workstation team at IBM Research: Randy Moulic, Nick Dono, Dan Dumarot, Cliff Pickover, Del Smith, and Dave Stevenson.

Furthermore, we would like to thank Michael Doggett and Michael Meißner of the Computer Graphics Lab at the University of Tübingen for proof readings and good suggestions to improve the text on parallel architectures and parallel programming.

For the parts on parallel volume rendering, Claudio is deeply indebted to A.E. Kaufman, J.S.B. Mitchell, C. Pavlakos, C.R. Ramakrishnan, who co-authored the original research he is briefly reporting in this tutorial. Most of this research was supported by CNPq-Brazil under a Ph.D. fellowship, by Sandia National Labs and the Department of Energy, and by the National Science Foundation (NSF) post-doctoral grant CDA-9626370.
Appendix A: Literature and Internet Resources on Parallel Programming

Books

- D. Butenhof - Programming with POSIX Threads, Addison-Wesley, 1997. This tutorial is based on this book. Besides being a good introduction into threading, it offers many details and knowledge of how to use threads.

Webpages

- www.lambdacs.com/FAQ.html - thread FAQ list
- www.best.com/ bos/threads-faq/ - thread FAQ list
- cseng.awl.com/bookdetail.qry?ISBN=0-201-63392-2&ptype=1482 - Additional information on Dave Butenhof book (e.g. source code)
- liinwww.ira.uka.de/bibliography/Os/threads.html - thread bibliography
- science.nas.nasa.gov/Software/NPB/ - results of parallel benchmarks
- www.sun.com/workshop/threads/ - SUN's thread pages
- www.mit.edu:8001/afs/sipb/user/proven/XMOSAIC/phthreads_man.html - pthread pages at MIT
- www.netlib.org/{pvm3,mpi} - a source for documents and packages for MPI and PVM.
- elib.zib.de - also a source for documents, packages, and more.
- www.erc.msstate.edu/mpj/ - MPI home page at Mississippi State.

Newsgroups

- comp.parallel - general newsgroup on parallel stuff
- comp.parallel.pvm - newsgroup on PVM
- comp.parallel.mpi - newsgroup on MPI
- comp.programming.threads - newsgroup on threading
- comp.sys.sgi.bugs - newsgroups for threading problems on SGI workstations
- comp.sys.sgi.{graphics, hardware, misc} - if the previous newsgroup does not help....

Appendix B: Tiny Thread Glossary

Barrier - All participating execution entities are synchronizing at a particular point within the parallel application. This point is called a barrier.

Cache-coherent - Modern processors use caches to speed-up memory access. On multi-processor systems this can result in different views of memory content for the individual threads. If a system is cache-coherent, special communication protocols ensure the same memory view. This system is called cache-coherent.

Concurrency - Parallel execution of programs. This parallel execution can be either time-sliced (on single processor machines), or really parallel on multi-processors.

Condition - A signaling mechanism to indicate the state of a shared resource.

Kernel thread - A kernel is a execution entity which is scheduled by the operating system kernel (one-to-one mapping).

Light-weight process - Physical scheduling entity of an operating system. On some systems, scheduled threads are mapped on light-weight-processes for execution. On Sun/Solaris systems, the kernel threads are called light-weight processes.

Message-Passing - Execution entities communicate by sending message exchanged via a interconnection network.

Mixed-model scheduling - Scheduling model inbetween user and kernel threads. Some scheduling tasks are performed by the thread library, some by the operating system kernel (many-to-few mapping).

Mutex - Synchronization mechanism for mutual exclusion of critical sections in a parallel program.

NUMA - Non-Uniform Memory Access - Main memory is distributed to the different hierarchy stages. Therefore, the memory access times are varying, depending on the processor and the physical location of the memory address.

Oversubscribing - More threads than processors are started. This is only efficiently possible with mixed-model scheduling or with user threads.

Preemption - A process, or a thread is disabled from execution (preempted), because the scheduling algorithm decided that another process/thread is more important than the current.

Process - An execution entity, containing a whole execution context (private address space, program counter, etc.)

Pthread - Thread standard.
Recycle thread - After performing its task, a new task is assigned to the thread; the thread is recycled.

Scheduling - Decisioin which execution entity can use particular resources (e.g. periphery devices, processors, etc).

Semaphore - Synchronization mechanism similar to mutexes. In contrast to binary mutexes, semaphores can have more than two states (they are “counting”).

Shared-memory Paradigm - Execution entities communicate via memory which is accessible from all entities.

Synchronization - If a shared resources is needed by different threads, their access must be handled consistantly. The threads need to agree on an order of access to the resource.

Thread - Control flow entity within a process. Threads of the same process share parts of the execution context (such as address space). Therefore, context switching, creation and destruction of a thread is much faster than for a process.

UMA - Uniform Memory Access - Access times to main memory are the same for all processors in a system.

User thread - Execution entity of a thread library. The library itself is scheduling the user thread (many-to-one mapping).

References


