

HP SEPIA System

A Scalable Visualization Cluster

See the future with SEPIA

System Overview

This document contains confidential technology of Hewlett-Packard Company or its affiliates. It is intended for use by HP customers or partners. No parts of this document are to be extracted or transferred to other parties without the consent of HP.

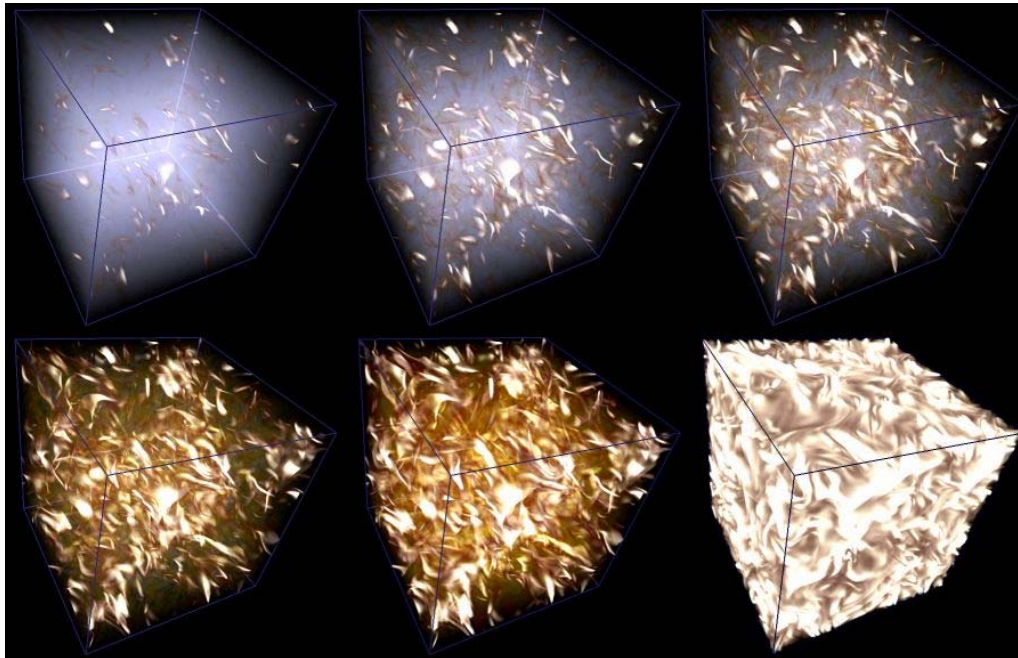


FIGURE 1. Interactive volume rendering of vorticity among a turbulent flow.
Image by Dr. Ravi Samtaney, Cal Tech University.

Introduction

HP has developed an advanced visualization system, namely, the HP SEPIA System. This innovative system leverages advances in clustering, graphics, and networking technology to provide scientists and engineers with the visualization and analytical power they need to solve some of the most challenging technical problems of the day.

The SEPIA System is based on HP technology that led to the award of a contract by ASC (Advanced Simulation and Computing) PathForward (a DOE National Nuclear Security Administration (NNSA) program) to leverage HP technology with commercial off-the-shelf systems for adding visualization capability to clusters of industry-standard systems. A very early version of the technology was first demonstrated at SuperComputing 2001.

This document explains the rationale for HP's system, its architecture, capabilities, components, and vision.

Today's technical challenges

As scientists and engineers tackle complex technical problems, they bring to bear the ever-more powerful computational resources available from scalable systems. This has overcome the hurdle of being able to process and generate the raw data in a timely fashion. However, it has created a new problem in that the large amount of data increasingly defies effective analysis.

Our ability to generate these large data sets... has completely outstripped our ability to visualize them, both for deriving science and verifying correctness.

Hugh Couchman
McMaster University

Here are some examples of technical challenges being tackled today:

- Norsk Hydro: How to simulate production from reservoirs of various hydrocarbon components of oil and gas with pressure variations. Analyze the whole reservoir rather than parts. Do history matching after production to improve future simulation.

Today's technical challenges

According to Western Geophysical, exploration and production for oil and gas can generate over 1 Petabyte of data per week.

- State University of New York, Stony Brook: Render interactively full-color Visible Female volume data of approximately 40 Gbytes.
- SHARCNET, Canada: Develop a volume rendering capability for astrophysical magneto hydrodynamic simulations that provides real-time interaction.

These examples and others span key industries that include Aerospace, Oil & Gas, Automotive, Manufacturing, Research & Defense, and Medical and Scientific Research. All these organizations share the need to reduce their costs of visualization by moving away from proprietary solutions with their inherently high fixed costs. These organizations also share the need to render complex 3D and 4D simulation data, using techniques such as volume rendering and isosurface extraction.

In many cases, the solutions to these technical challenges generate increasingly large data sets that can be on the order of 1 Terabyte to 1 Petabyte. Scientists and engineers need a visualization platform that not only supports visualizing such data statically, but supports real-time interactivity.

The purpose of computing is insight — not numbers.
R.W. Hamming

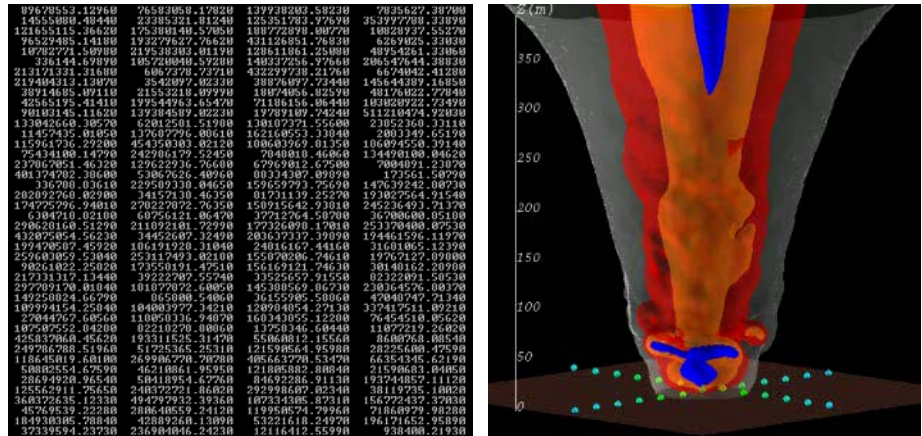


FIGURE 2. Wind direction and isosurfaces to show pressure simulations from the interaction of a tornado vortex with the surface.
Image by A.Gel, W. Virginia University and G.Giras, Pittsburgh Supercomputing Center.

Higher resolution display systems are needed to support output devices such as stereo graphic walls and CAVEs, which display about 6 Mpixels. The display requirements can extend to 100 Mpixels and beyond; single LCD monitors can now generate 9 Mpixels.

Today's visualization solutions: costly, inadequate

Often, the goal of such large displays is the collaborative analysis by teams of scientists and engineers in problem-solving environments. This is the outgrowth of a trend in which scientists and engineers located around the world actively collaborate on solving their technical problems. To do this effectively, they need to visualize and analyze their data on visualization systems that are powerful and flexible in terms of application and display.

Today's visualization solutions: costly, inadequate

Proprietary scalable visualization solutions are characterized by high costs and inflexibility given today's rapid innovation in relatively low-cost high-performance graphics adaptors for the PC gaming markets.

Most harmful to the long-term development of any proprietary system is its inability to take advantage of the rapid innovations available from off-the-shelf components, most notably, processors, graphics adaptors, and high-speed interconnect networks. Finally, proprietary systems are often tied to proprietary operating systems rather than the increasingly popular and dynamic Linux operating system.

Competing products also lack true volume rendering or support only limited volume rendering and perspective views without image rotation, all of which requires node reordering and blending semi-transparent images. They also lack support for allocating cluster resources to multiple users.

HP SEPIA System

The HP SEPIA System is a scalable visualization solution that brings the power of parallel computing to bear on your most demanding visualization challenges.

The SEPIA System leverages the advances made across the industry in PC class systems, graphics technology, processors, and networks by integrating the latest generations of these components into its clustering architecture. It combines these off-the-shelf components with its own image compositing sub-system based on Field Programmable Gate Arrays (FPGA). This unique base of hardware and associated firmware underlies Linux clustering software and is further enhanced by a set of software libraries developed by HP and its partners to facilitate the use of the system by new and existing user applications.

SYSTEM VIEW

In HP's vision of SEPIA, it integrates seamlessly into the complete computational, storage, and display environment of customers as shown in Figure 3. This vision is uniquely possible due to HP's leadership position and extensive experience in the high performance technical computing arena.

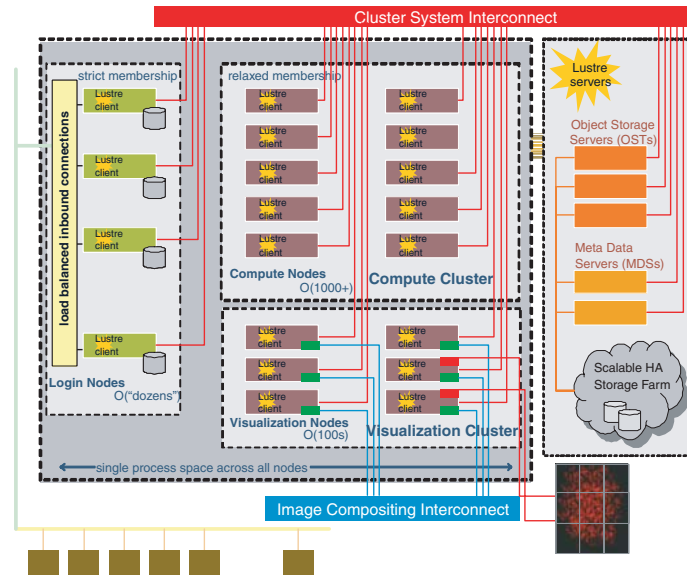


FIGURE 3. System view of a computing environment with integrated SEPIA System.

High-speed interconnect networks have made feasible the transfer of large amounts of data among the following: individual users at their desktops or logged into cluster; servers that are part of data storage farms; the compute cluster; the visualization cluster; and local and remote display devices.

A typical usage model for the type of system shown in Figure 3 involves the following:

- A compute intensive application, for example, automobile crash test simulation, runs on the super computing compute cluster of the environment.
- The large data set generated can be stored in the storage servers for later retrieval or directed in real-time for rendering on the SEPIA cluster portion of the system.
- One or more users can log their sessions concurrently into the SEPIA System, which can allocate resources efficiently to meet the rendering and display requirements of each user's application.

SEPIA architecture

- Users' visualization applications use parallelization techniques to distribute their graphical rendering across the SEPIA nodes, each of which in turn renders a portion of the output for the final image.
- Image data flows through a chain of rendering nodes at interactive frame rates over SEPIA's high speed network. As the data moves down the chain, each node "assembles" its piece of rendered image data with that of an upstream node.
- At the end of the rendering chain, a fully composited image is available for display locally or remotely.

The SEPIA System serves as a key unit in an integrated computing environment whose end result is to display the results of the generated data in those locations where scientists and engineers can most effectively carry out their analyses either individually or in collaboration.

SEPIA architecture

The SEPIA System derives its most powerful attributes from its architectural design consisting of a cluster of rendering nodes, network-based compositing using sort-last techniques, and flexible graphical compositing operators.

Its image-based approach works with a variety of visualization techniques, including isosurface extraction and volume visualization. Such a graphics architecture combines the high performance of clusters of rendering machines and the interactivity made possible by the speed, scalability, and low latency of its compositing network.

CLUSTER MAKEUP

HP's SEPIA — through the use of its powerful compositing technology — offers a graphics visualization solution that can be used by a variety of applications that can run on distributed computing systems, in this case, a cluster of Linux workstations. Figure 4 illustrates the makeup of the cluster:

- Industry standard compute nodes with standard OpenGL 3D graphics adaptors serve as render nodes and run clustering software and Linux. Each node also has an HP-designed PCI-X image compositing sub-system. Use of off-the-shelf graphics adaptors lets the system take advantage of new generations of adaptors as they are available.
- A Windows or Linux master node runs the application and possibly the administrative functions of the cluster. The choice of two operating systems on the master node provides more options for the types of applications that can run on the SEPIA System.

SEPIA architecture

- Display nodes transfer DVI or RGB output to the display device and synchronize multi-tile displays. A range of displays are supported, including large powerwalls, immersive CAVE displays, and locations local and remote to the SEPIA System.
- The high speed image network infrastructure relies on the latest InfiniBand fully duplexed, 4x standard. The image compositing sub-systems are connected using this low latency switched network using a Clos non-blocking switch, which routes the data among the InfiniBand ports. The use of this low latency switched network makes possible the interactive rendering so important to visualizing large data sets.
- The Gig/E cluster network supports cluster management. High-speed low-latency interconnects such as Myrinet, Quadrix, or InfiniBand could be added.

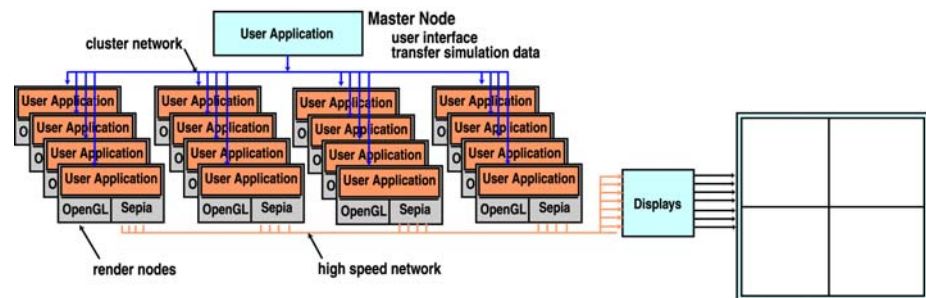


FIGURE 4. SEPIA pipeline components.

The network-based image compositing sub-system implements a sort-last architecture for compositing a 2-D or 3-D image. The compositing sub-system acquires images from standard graphics adaptors through their DVI-output. The compositing sub-system acquires image data including RGB, depth, alpha, and stencil with minimal impact on the rendering process. Compositing takes place within the compositing sub-system in each render node, and the image data flow through chains of compositing sub-systems over the high speed network without having to be read back into system memory — a key driver of enhanced performance.

The final images are output through the graphics card in display nodes. You can synchronize the output of compositing sub-systems to avoid visual artifacts and support multi-tile active-stereo displays.

Final images can also be read back into memory for transmission to a remote workstation display over a network external to the cluster. This lets users interact with applications running on the cluster from their offices. Images read back into system memory are available to applications. For example, an application could

SEPIA System attributes

choose to write the images to files. The SEPIA System can also acquire image data from system memory to accommodate special rendering needs, for example, PCI cards that are specialized for volume rendering.

Figure 4 also shows a master node that is communicating with the render nodes over the cluster network (as opposed to the high speed network used for image compositing traffic). The cluster network carries file I/O and application communications, for example, MPI traffic. The user interface for a visualization application runs on a master node and communicates with the render nodes over the cluster network, sending control information such as changes in point-of-view or sending data or OpenGL.

DISPLAYS

SEPIA Systems support a wide range of displays and configurations: from single displays to tiled displays in walls and immersive CAVE environments with active/passive stereo. Depending on the demands of your display devices, you can take advantage of DVI, RGB, or video output. The aggregate resolution of these displays can range from 10's to 100's of megapixels.

SEPIA System attributes

SCALABILITY

The key to the SEPIA System's scalability and flexibility is its compositing technology. This technology couples render nodes together into one or more chains through a switchable high-speed display subsystem that routes images and performs image compositing operations at real-time rates. The compositing system allows for scaling up the number of nodes working on a problem in parallel to handle larger data set sizes, to increase frame rates, and to display at higher image resolutions.

The characteristics of the SEPIA System that support effective scaling fall into several areas:

- Performance scaling results from having portions of the image data rendered on separate nodes in the SEPIA System. In effect, the work is divided up among the nodes working in parallel. Larger data sets can be accommodated by more render nodes. The system design can scale from 8 to 1,000 or more render nodes.

The parallel nature of the rendering chain removes a key performance bottleneck of a conventional hardware-accelerated graphics architecture, which feeds data sequentially to a centralized pipeline.

A key feature of the image compositing sub-system that makes such scaling possible is its ability to composite image data that flows directly from one render node to another. Once an image is constructed by a graphics adaptor, it

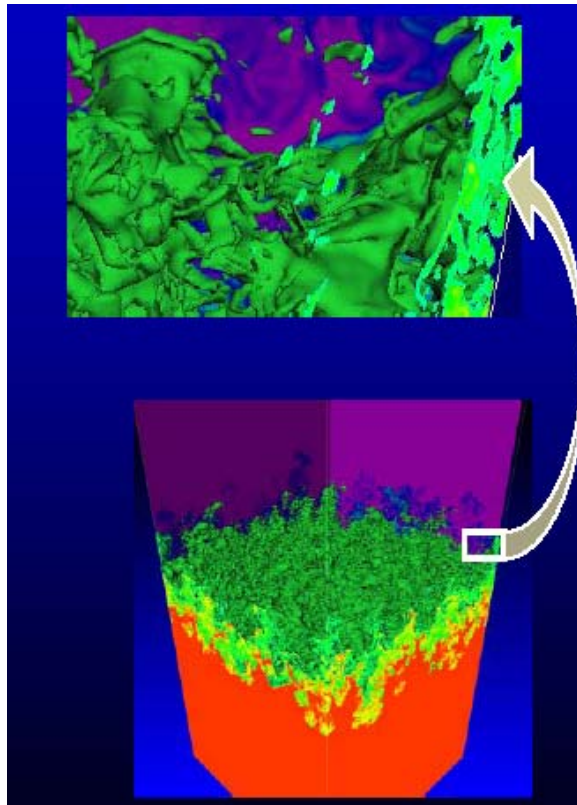
travels over a high-speed interconnect and data paths, and goes through the compositing sub-systems but never goes back into system memory or a graphics adaptor. This is a significant driver of improved performance.

In addition, the high-speed InfiniBand interconnect transmits data among the image compositing sub-systems with adequately low latency and high-speed to maintain interactive frame rates for delivery to the display node. Use of the InfiniBand standard by SEPIA has the additional advantage that it is an inherently scalable interconnect intended for large clusters. Furthermore, its components are available from multiple vendors and this leads to cost savings.

- Resolution scaling results from the SEPIA System's ability to perform parallel rendering combined with the parallel display of multiple tiles. This capability makes it possible to display high resolution data and to display to large display surfaces, including immersive displays and display walls.

FIGURE 5. These images illustrate the demands for extremely high resolution displays, which are the result of visualizing extremely large data sets, for example, many over 10 TB. Resolution scaling such as that made possible by the SEPIA System is needed for this type of visualization challenge.

Image courtesy of Lawrence Livermore National Laboratory.



FLEXIBILITY

One of the most powerful attributes of the SEPIA System is its flexibility, which makes it possible to apply it effectively to a wide range of technical problems. This flexibility derives from several architectural characteristics of the system:

- Use of Field Programmable Gate Array (FPGA) technology: Its use within the image compositing sub-system makes it possible to create programmable compositing operators in the firmware that can accommodate various pixel formats and application requirements that may change over time.

A major fraction of the data contained in the large datasets is never directly observed by scientists. Scalable, interactive visualization systems are a key enabling technology making it possible for scientists to perform detailed exploration of a greater fraction of the data than previously practical.

Lawrence Livermore National Laboratory

- Network-based compositing: This key capability supports non-commutative compositing operators that require partial images to be composited in an order that is view-dependent. This is essential for distributing polygonal and volumetric data that is rendered and composited as semi-transparent images. For example, it makes it possible to divide a volume of data into subvolumes that are assigned to render nodes without replicating data. As the view of the volume is rotated, the order of the nodes in the compositing pipeline can be changed to preserve the correct compositing order; there is no need to change the assignment of subvolumes to render nodes.
- Network-based compositing is essential for interacting with composited images with semi-transparent polygon data; it provides support for fragment processing and complex rendering modes, such as volume rendering. It makes it possible to divide a volume of data into more than one subvolume. Because blending subvolumes is non-commutative, dynamic reordering of the rendering nodes (a characteristic of network-based compositing) is necessary in order to avoid replicating data from one rendering node to another in the pipeline. For example, a cube divided into several slabs would need to have the rendering order of the slabs changed as the cube is rotated to preserve the proper transparency among the slabs.
Network-based compositing also makes it possible to dynamically load balance an application's work. For example, using a SEPIA System, you can change the number of nodes assigned to a given tile display at run-time to accommodate a disproportionate share of work needed for that tile.

In a related way, use of the network to connect the render nodes makes it possible to assign render nodes flexibly under user and software control to meet the needs of multiple users of the cluster and the demands of an application. Hard-wired cabling configurations preclude this flexibility.

- **Image data routing:** The SEPIA System can route image data dynamically from a node to the display. For example, if three projectors are used to display a volumetric image, interactively turning the image causes portions of the image to occupy other areas of the display – areas covered by a different projector. Image data routing makes it possible to send the same image data to a different projector or split it without the need to replicate the model data. This can produce significant savings in model data storage, reduce the number of nodes needed in the system, and ultimately make volumetric image manipulation possible at interactive rates.
- **Dynamic node configuration:** Changes in your application or display requirements may result in a desire to reconfigure the SEPIA System's allotment of render nodes, display nodes, display devices, and the number of nodes assigned to a display tile and device. The SEPIA System design lets you reconfigure node assignment and pipeline assignment via software and without recabling. All the render nodes of a large SEPIA System can be used by a single job to display a large data set on a huge multi-tiled display. Moments later, you can divide the same render nodes among a number of smaller jobs, completely under user and software control without recabling. Thus, the Sepia System is designed as a true cluster environment in which you can allocate resources as needed.

Finally, when these architectural characteristics of the SEPIA System are integrated with an HP high performance compute cluster (see Figure 3), you can select an optimal number of application or compute nodes and match these with an appropriate number of render nodes. Visual applications with high computation requirements can be distributed over the compute nodes and the visualization render nodes, thus the render nodes can double as compute nodes.

This flexibility is critical because often visualization applications need to perform intensive computations to compute isosurfaces, streamlines, or particle traces. You can select the application nodes based on factors such as model size, and match those to the rendering nodes your application needs to yield the desired rendering performance and resolution.

Compositing

SEPIA System compositing operators are implemented in a programmable FPGA. Supported operations include depth compositing, alpha blending, antialiasing, and spatial compositing.

Applications differ in terms of the best way to carry out their image compositing. With this in mind, the SEPIA System supports many approaches to partitioning the compositing work — including support for combining the approaches — with the intent of making the job of application development easier and more effective.

POLYGON DATA PARTITIONING

Data partitioning provides linear scaling of the data set as the number of render nodes increases — with negligible reduction in performance.

You can partition models consisting of millions to hundreds of millions of polygons across a set of render nodes and render in parallel to improve the frame rate and reduce latency. SEPIA depth compositing uses the color and depth data to combine the partial images from the individual render nodes into an image for the complete model. Since the model data is not replicated on each node, huge datasets can be accommodated by the aggregate system and graphics memory of the render nodes.

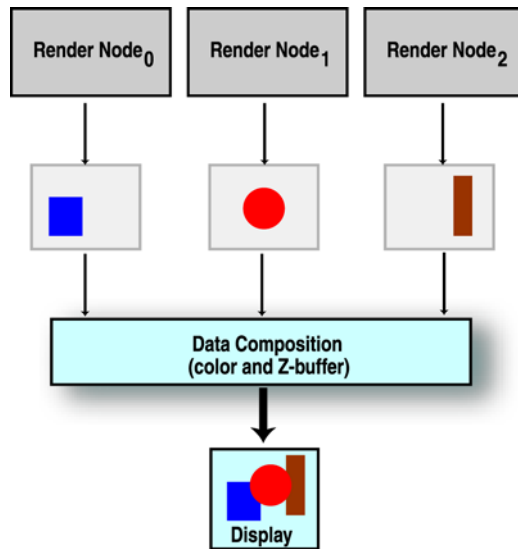


FIGURE 6. Data partitioning scheme used for compositing.

VOLUMETRIC DATA
PARTITIONING

Volumetric data partitioning allows for interactive viewing of datasets many times larger than is feasible on a single workstation.

You can subdivide volumetric data among render nodes without replication and the subvolumes can be rendered independently to produce an image of the subvolumes. These images can include transparency, for example, to show layered surfaces or to show different material densities.

SEPIA's alpha blending compositing uses the color and alpha components of the pixels to combine the images. Blending requires that the images combine in the proper order. As the view of the data changes, for example, by rotating the volume, the compositing order must change. Because SEPIA routes images through a switched network, the compositing order of the subvolumes can be changed between frames without any impact on the interactive performance.

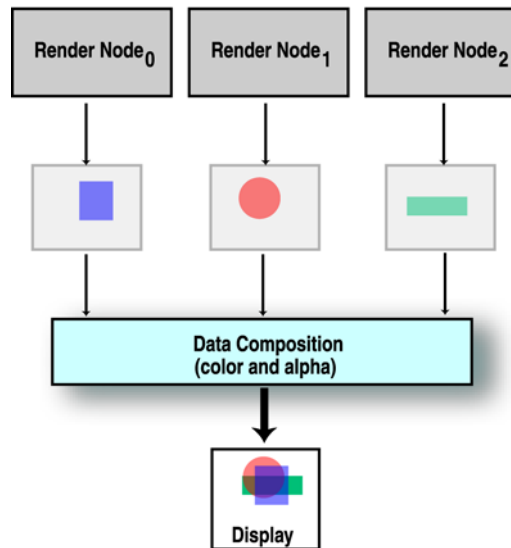


FIGURE 7. Volumetric data partitioning scheme used for compositing.

FRAME-INTERLEAVED
RENDERING

For applications where interactive latency and dataset size is not an issue, frame-interleaved rendering is a simple approach for parallelizing an application to improve frame rate.

In this approach, different rendering nodes render consecutive frames in parallel. The results are routed in sequence to a single display. Unlike the previous data partitioning approach, this approach does not decrease the latency of rendering.

Also, the model must be replicated on each render node. However, it is easy to divide complex rendering this way; it scales frame rate linearly with the number of render nodes; and it can increase frame rate when combined with other partitioning techniques. It can also be used to generate stereo pairs.

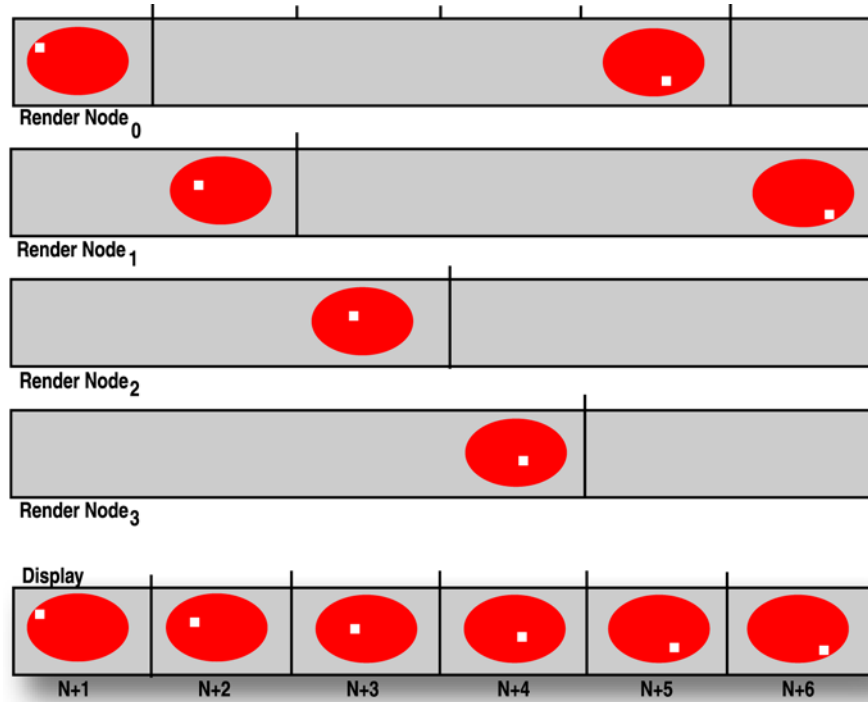


FIGURE 8. Frame interleaved rendering scheme (temporal scheme) used for compositing.

SPATIAL IMAGE PARTITIONING *Spatial image partitioning provides a simple mechanism for improving frame rate in applications that are pixel-fill limited.*

The overall image for a single frame can be partitioned spatially into subimages that are rendered in parallel on a set of nodes. SEPIA compositing can be used to stitch the subimages back together into a single image. Like frame-interleaved rendering, it is easy to divide the rendering, since the subimage can be produced by exactly the same rendering steps as the overall image with only a change to the viewing frustum. It is difficult to avoid data-replication in this approach and so it does not lend itself to large datasets.

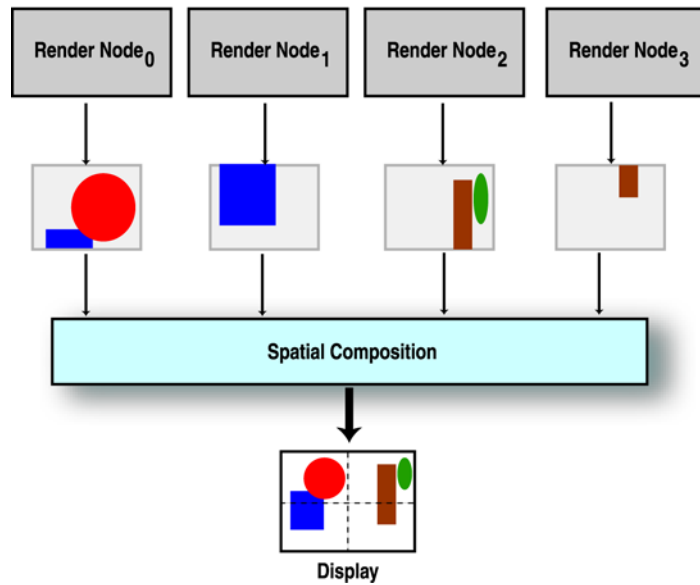


FIGURE 9. Spatial partitioning scheme used for compositing.

JITTERED SAMPLE PARTITIONING

Jittered sample partitioning improves the image quality by diminishing jagged edges and smoothing lines.

Full screen antialiasing (up to 16X) can be implemented by averaging several jittered samples of a scene. SEPIA allows samples to be rendered in parallel and combined using SEPIA's antialiasing operator for improved rendering performance. These samples can include antialiasing implemented on the graphics adaptors for even greater image quality.

Application support

HP recognizes that a key capability of the SEPIA System is to make it possible for applications to run without extensive re-coding. To that end, HP is working with both commercial ISVs and the Open Source community to ensure solutions are available for SEPIA. Current Open Source projects are underway for interests such as microscopy and imaging research, large-scale visualizations, large dataset visualization, magneto-hydrodynamics, parallel distributed rendering, and others.

Figure 10 illustrates the layers of software support that are part of the SEPIA System.

- Cluster management software and visualization resource management software.
- APIs needed by visualization applications that provide access to information about the visualization resources and provide access to the compositing features and connected displays.
- Visualization toolkits and libraries.

Visualization and Graphics toolkits are provided by third party vendors and the Open Source community. ISV applications and applications written by end-users can run on SEPIA Systems taking full advantage of the various toolkits, libraries, and a rich set of APIs. SEPIA Systems use+ standards such as OpenGL, Linux, InfiniBand, and DVI for portability and interoperability.

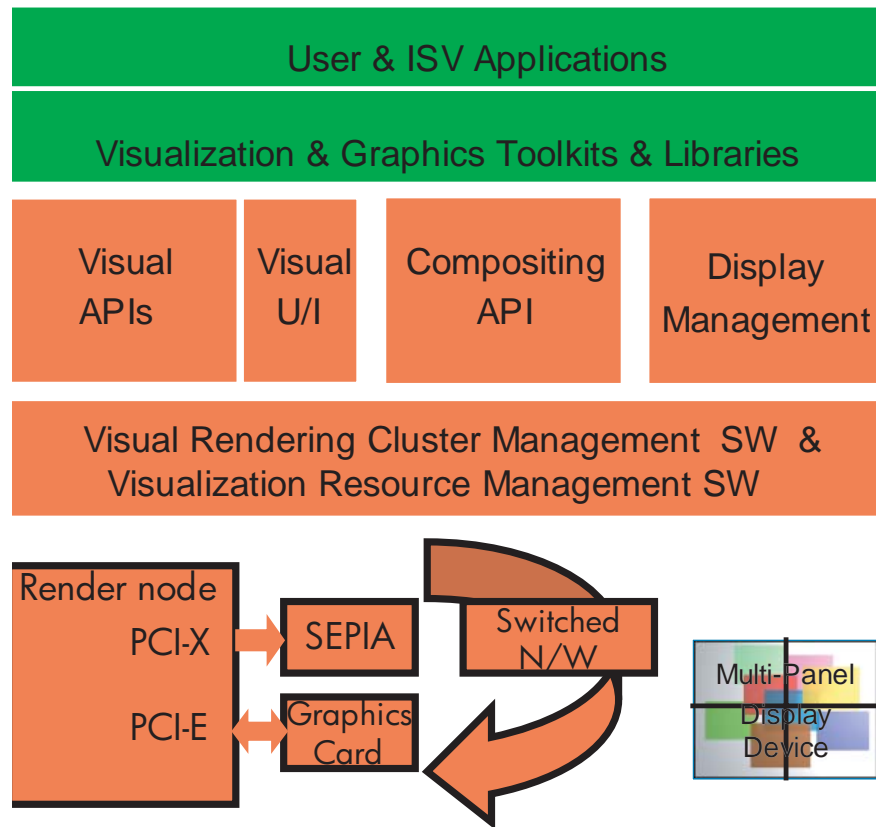


FIGURE 10. Software support for application development and use.

Application support

To run on a SEPIA System, an application must be parallelized and distributed. There are two main pathways to get to this state: OpenGL applications made parallel by design and serial applications made parallel automatically through middleware libraries or toolkits, for example, scene graph applications.

OPEN GL APPLICATIONS

If your application is already a parallel and distributed one, you can use OpenGL directly. You need to make a limited number of calls to the System Compositing API. These calls do some setup work, including selecting compositing operators and synchronizing the work among the render nodes.

Most visualization applications already support OpenGL directly or through graphics toolkits. Auto-parallelizing toolkits, such as Chromium, allow standard OpenGL applications to run on SEPIA Systems, although without the full performance advantages of a true parallel application.

SCENE GRAPH APPLICATIONS

SEPIA Systems let you take advantage of enhanced support for scene graph applications available through scene graph middleware libraries and toolkits compatible with the SEPIA System Compositing Network API. The result is that the application quickly can be made available on the SEPIA System and take advantage of its parallel scalability features.

COMPOSITING NETWORK API

The Compositing Network API for the SEPIA System lets you access the functions of the image compositing sub-system. This API provides such key functions as setting up the environment, dealing with properties, starting operations, stopping operations and so on.

Much of the functionality is performed by hardware under the direct control of firmware. As part of setting up your system, you download the firmware, which allows for additional development and customizing.

Key operations available via the Compositing Network API include:

- Adding sort-last compositing to your application with the addition of just a few API calls.
- Using any number of nodes in your cluster to produce the rendering results that you need.
- Assigning your render and display nodes in virtually any way to solve your problem or combining them to put more power into your calculations. For example, one day, you might configure a twelve node cluster to have eleven render nodes and one display. The next day, you re-assign the nodes such that there are two pipelines with four renders nodes each and two display nodes each.

Key benefits

- Dynamically reconfiguring your render nodes while your application is running. This lets you use your cluster to solve your problem in the most efficient way for your application — and tune during your run.
- Choosing compositing operators and setting parameters for the operators.

The Compositing Network API provides the application developer with a toolkit of routines specifically customized for use by graphics applications running on the SEPIA System.

Key benefits

- Network-based pixel compositing engine provides power, performance, and scalability with node reordering for full volume rendering.
- A sort-last compositing architecture drives scalability and performance.
- A true cluster environment with variable resource management and assignment supports multiple users and varied application demands.
- Commodity off-the-shelf components, namely, processors, graphics adaptors, interconnects, and Linux, preserve your investments and enhance price-performance ratios.
- FPGA-based programmable pixel operators provide flexibility and extensibility in that they support a solid base of initial visualization requirements with the promise of additional operators as needed.

For more information

Contact your HP Sales representative for more information on SEPIA Systems.

You can also access the following website:

HP Visualization Collaboration Competency
www.hp.com/techservers/hpccn/

For more information

© Copyright 2004,5 Hewlett-Packard Development Company, L.P.

The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein. Microsoft and Windows are registered trademarks of Microsoft Corporation in the United States and other countries. Xeon and Itanium are trademarks or registered trademarks of the Intel Corporation. All other product names mentioned herein may be trademarks of their respective companies.

Printed in the U.S.A. March, 2005

