HRG Assessment:

Cassatt Collage Run-Time Fabric:
An Application Execution Environment

Cassatt Collage™ run-time fabric provides an enterprise-class application execution environment that allows an IT manager to view a collection of commodity servers as a single, seamless expanse of fabric on which to run applications. Collage provides the capability to manage the fabric with tools to easily boot large numbers of servers, deliver a consistent version of the operating system across servers, monitor the health of servers, handle alarms, and perform other management tasks. Cassatt Collage addresses one of the key challenges of IT departments today — how to realize the benefits of low cost commodity hardware while addressing the increased complexity of managing large numbers of servers.

Cassatt Collage today is focused on delivering value to IT organizations with High Performance Computing (HPC) workloads in technical or business applications. In addition to traditional High Performance Technical Computing (HPTC) applications, Collage is suitable for High Performance Business Computing (HPBC) applications such as data mining, data modeling, data warehousing, and other analytics. These applications are characterized as compute intensive, I/O intensive, and parallel in nature. Future releases of the Collage run-time fabric will deliver an application execution environment more suitable for commercial, transaction-oriented, data intensive workloads.

In the longer term, Collage will evolve to an application execution environment ("an operating system for the enterprise") that supports the industry direction towards on-demand computing and a broader range of enterprise applications, including service-oriented architecture (SOA)-based applications. On-demand computing provides an IT environment where resources are available when needed, and a run-time fabric, such as Cassatt Collage, will be required to deliver the flexible allocation of resources/workloads required by on-demand computing. SOA, with its modular approach to applications, facilitates allocation of workloads. Delivering this application execution environment is key to Cassatt's approach to on-demand computing to improve flexibility and reduce cost/complexity.

Underpinning Collage's application execution environment is an architecture based on sophisticated role-based computing. The majority of nodes in a Collage fabric are application nodes, meaning that their main role or responsibility is to run applications. Service nodes provide system services. They are responsible for administration and various types of system services such as I/O. Role-based computing allows system services to be assigned to specific hardware nodes in a Collage run-time fabric. The primary benefit of roles is that they provide the flexibility to scale a system by allowing more resources, of the correct type, to be added when and where demand dictates. Previous attempts to build a run-time fabric have failed because the run-time systems infrastructure was not flexible and could not scale.

In this paper, we outline the current market trends, present the business value that Cassatt provides to customers, characterize Cassatt's product offering, and generally describe Cassatt's approach to providing value to its markets.

Harvard Research Group, Inc.
P.O. Box 297
Harvard, MA 01451 - 0297 USA

Tel. (978) 263-3399
Fax. (978) 263-0033
hrg@hrgresearch.com

Market Trends

Many technologies adopted by commercial IT departments had their origins in scientific and technical computing environments: UNIX, RISC processor architecture, HTML and the World Wide Web, and Linux, among others. Today, HPC in the technical space is contributing to innovation in the enterprise. Grid computing (both hardware and software aspects) and Linux clusters are two good examples of HPC technologies, which are making the transition to enterprise computing. Commercial applications such as those in financial services are now using these technologies to lower cost and improve performance, solving problems in several minutes that used to take many hours.

While both grid computing and Linux clusters have demonstrated price/performance benefits from a hardware perspective, the cost and challenge of managing these environments is one of the key obstacles to commercial adoption. The complexity of managing ever-increasing numbers of commodity servers is significantly more difficult than the management of a smaller number of high cost, high performance servers.

Value Proposition

Cassatt targets IT organizations responsible for enterprise infrastructure and datacenter management. The Collage run-time fabric is an excellent platform for managing large numbers of commodity servers because it has the flexibility to serve as an application execution environment for a broad range of workloads while providing a single unified image to users and system administrators.

Collage can deliver value to customers who have computationally intensive, batch oriented critical applications and an interest in using technology to maintain competitive advantage. Those customers who buy significant amounts of compute power to meet the demand of peak processing periods — such as in the financial services market — and need significantly less capacity during non-peak periods are also potential customers. Collage presents these enterprises with an opportunity to provide peak-level services while leveraging trends in commoditization and standardization to deal with capacity management and providing lower total cost of ownership (TCO).

Value Proposition: Determining the Value

The Collage run-time fabric delivers value in several areas by:

- Reducing management costs with an enterprise cluster/grid management solution
- Improving price/performance by enabling the transition from expensive proprietary systems to an application execution environment built on commodity servers
- Providing flexibility to meet datacenter needs with re-allocation of server roles in a Collage run-time fabric as needs dictate

With its professional services organization, Cassatt can also provide assistance in developing a roadmap to lead an organization toward a service-oriented infrastructure — an application execution environment that supports a service-oriented architecture and services-based applications.

It is HRG's perspective that Cassatt's vision and product plans will help enterprises move toward the goal of on-demand computing and reductions in cost and complexity for IT.

Overview of Collage

Collage delivers an application execution environment that looks to the user like a single unified system. Collage supports a workload mix that can consist of commercial applications, such as data mining applications, compute-intensive applications that require Message Passing Interface (MPI) and fast interconnects, and a variety of other workloads, some of which may be Web services-based applications.

Tel. (978) 263-3399
(978) 263-0033
hrg@hrgresearch.com

The basic goals for Collage are to:

1. Provide an application execution environment capable of managing large and dynamic workloads on systems that have thousands of processors, mixed workloads including interactive jobs, large-scale parallel applications, and I/O intensive jobs.

2. Provide a single unified system view for users and applications.

3. Provide a single point of management for systems administrators.

4. Maintain application programming interface (API) compatibility.

5. Support the leading server operating systems: Linux (today) and Windows (future).

6. Support industry standard processors, such as Intel 32-bit processors today and in the future 64-bit processors.

Collage is a layer of software on top of a cluster/grid of potentially hundreds or thousands of nodes. It virtualizes this collection of physical nodes, making them appear as a single large-scale symmetric multiprocessing (SMP) system for the purposes of system administration, application execution, scheduling, system utilization, resource sharing, and global I/O. However, it is a distributed architecture consisting of physical nodes, with each node having its own memory system.

The underlying infrastructure on which Collage software runs consists of a cluster of server nodes. These nodes are connected via Gigabit high performance interconnects. Collage uses the concept of roles (see Table 1) to create a hierarchy of application and service nodes.

The upcoming release of Collage provides a Web-based portal for administrative control and makes improvements to reliability, availability, and serviceability (RAS) of the run-time fabric and performance by removing most single points of failure. It offers an application execution environment for workloads typically described as HPBC tasks, including data mining, data warehousing, and compute intensive crossover applications, as well as traditional HPTC.

Collage is a suitable application execution environment for one or more of the tiers in an n-tier architecture. That is, some nodes in a Collage cluster may be configured to run front-end software, some nodes could be application server nodes, and other nodes could host back-end databases. This is of value to a system administrator because the nodes in the application execution environment are managed using the same management solution.

Collage fits into an enterprise's current IT infrastructure. It is designed to meet the cluster/grid management solution needs of enterprise customers. To protect the investments of enterprise customers who already have high-level, enterprise-class management solutions such as CA's Unicenter, HP's OpenView, and similar management products from BMC and Tivoli in place, Collage can be integrated through SNMP.

Case Study: Collage Makes Software Builds More Cost Effective

Cassatt Collage augments the Linux operating system running on command and application nodes via its Unified System Services Layer (USSL). USSL provides a set of single-system capabilities that permit large distributed applications such as software builds (*parallel make*) to run in a cluster of small servers as they would in a large SMP system.

The software build of a large application consists of a series of compilations of individual source files into object code. These object files are linked into libraries, which are then combined in various ways to produce a set of

Tel. (978) 263-3399
(978) 263-0033
hrg@hrgresearch.com

executable programs. Individual compile steps rarely depend on results from prior compiles. A parallel make takes advantage of this dependency knowledge to schedule many compile steps to run at one time. Depending on the number of parallel steps and the processors available to the build, parallel make can reduce build times dramatically, by as much as an order of magnitude on a 12-processor system.

Parallel make often depends on the compile environment existing in the same server, which can limit the number of processors available. Collage software provides the single-system view required by a parallel make while allowing the seamless assignment of compiles to multiple application nodes, offering the same performance as an SMP server on low cost commodity hardware. This ability to use commodity hardware provides for dramatic improvements in price/performance. The single-system view of Collage allows the use of parallel make without the extensive cost and effort associated with developing custom scripts required by other clustering approaches.

In addition, increasing the processing power of a Collage run-time fabric is a simple matter of adding more application nodes as needed. In contrast, scaling an SMP server beyond the limits of a single box would require replacement with a larger and far more expensive server.

Role-Based Computing — Making Scalability Work

The underlying architecture for the Collage run-time fabric is based on *roles* (also referred to as node-specialization). Some server nodes are responsible for running applications while other server nodes, referred to as service nodes, are responsible for running management software, performing I/O or network functions, and providing resource manager functions. A node may be dedicated to a specific role, or a single node may have multiple roles. In larger configurations, several nodes may be responsible for the same role. This latter case typically occurs when there is a need for system services to scale as the run-time fabric grows.

Collage roles are important to customers for the following two reasons:

1. They allow a cluster to grow (scale) without the bottlenecks that occur with most other cluster management solutions.

2. They give system administrators the ability to re-provision the nodes in a cluster to meet the needs of various workloads — for example, the workloads run during normal working hours may differ from the workloads run during non-working hours, and they may require a different mix of service and application nodes. The number and assignment of roles to nodes is defined at boot time. In the future, Collage will permit dynamic re-deployment of roles between application and service nodes.

Scalability is a major challenge for most other clustering solutions. As nodes are added to meet customer demands, the overhead associated with managing the cluster increases, creating bottlenecks that can limit scalability. Collage prevents bottlenecks by allowing the system administrator to easily add more nodes to perform specific administrative and operating system service roles. This spreads the administrative and services overhead across multiple nodes, reducing the possibility of bottlenecks.

The flexibility provided by roles allows each cluster to be configured based on the types of workloads that it must accommodate. For instance, an I/O intensive application may need several dedicated nodes with I/O responsibility. Another environment, such as one in support of interactive applications, may require multiple command nodes for logins/commands.

Activities/services that are not of direct benefit to applications are moved from application nodes to service nodes. Service nodes generally support one or more of the six non-application roles in Table 1. They provide support for tasks such as user logins, job submission and monitoring, I/O, and administrative functions. The purpose of the service nodes is to improve throughput and system utilization. Table 1 summarizes the seven roles currently supported by Collage.

Tel. (978) 263-3399
(978) 263-0033
hrg@hrgresearch.com

Table 1: Collage Roles

Role	Description	Default Configuration Requirements
Admin	Provides single point of administrative access for system booting and system control and monitoring.	One admin role must be configured on each cluster. With the exception of the command role, this role can be combined with other service roles on a node.
Application	Runs user applications launched from a command node. Nodes assigned an application role cannot be simultaneously assigned another role.	Number of application nodes is determined by customer requirements. The current release of Collage supports systems with 8 to 512 application nodes.
Command	Provides for user login, application builds, submission, and monitoring.	Number of command roles is determined by customer requirements. At least one command role must be configured in the system. With the exception of the admin role, this role can be combined with other service roles on a node.
File System I/O and Network I/O	Provides for support and management of file systems and disks. This role can be used to support NFS from outside the system, or project local file systems to the rest of the system using Cassatt's Unified Parallel File Service (UPFS).	Use of I/O roles is optional. When assigned, the number depends on customer requirements. An I/O role can be combined with other service roles on a node; however, a node cannot be assigned both the file system I/O and network I/O roles.
Leader	Facilitates scaling of the shared root file system. Offloads network traffic from the node with the admin role. Required for large configurations.	Each group of nodes requires a leader to monitor and manage the group. A group can be up to 64 nodes. This role can be combined with other service roles on a node. It is recommended that systems with more than 16 application nodes be configured into multiple groups.
Network Director	Defines the primary gateway node on the system, which handles inbound traffic for all nodes and outbound traffic for those nodes with no external connections.	One network director role must be configured on each cluster. This role can reside with other service roles on a node.
Resource Manager	Defines the location of the system resource manager, which is responsible for allocating processors to user applications.	One resource manager role must be configured on each cluster. This role can be combined with other service roles on a node.

Source: Cassatt, June 2004

Tel. (978) 263-3399
(978) 263-0033
hrg@hrgresearch.com

Application Role

Applications run on nodes that provide an efficient execution environment while other nodes provide systems services. As the application requires additional compute resources, either by spawning a new process or through an MPI call, more application nodes may be recruited to provide additional resources. All application nodes have access to a shared file system projection that appears local to the node. The actual file systems are either on service nodes within the run-time fabric, or they are outside Collage. Application nodes can make use of local disks for swap space and temporary files, but they do not provide file service to other nodes within the system.

Administration Roles

As Table 1 indicates, service roles can be grouped into two role categories: *administration roles* and *operating system services roles*. The admin and leader roles are the administration roles. While the number of leaders is dependent on the number of groups of application nodes in the cluster, only one admin node can be configured on a system. These two roles provide the environment where system administration tasks are executed. They are not visible to end-users.

A node assigned an administration role is responsible for booting, hardware and services health monitoring, and other server administrative tasks. These tasks are shared across the admin and leader roles. The admin role provides a single point of access for system administrators. The admin node is connected to all other service and application nodes over a private network referred to as the administrative network (see Figure 1). The admin node contains the shared copy of the system software that is provisioned to the other nodes at boot time. In larger configurations, leader nodes act as cache servers for system software, reducing the demand on the admin node's file system.

Operating System Services Roles

Operating system services roles include the resource manager, network director, I/O, and command roles. These roles support users and applications either directly or indirectly. Only one instance of the network director and resource manager roles can be active, although second instances are supported for failover. A number of command roles will be configured such that the user login and application launch workloads are adequately handled. I/O nodes can be configured as necessary to meet the file I/O requirements of the cluster.

Users log into command nodes, where they are able to run and monitor applications. The network director node is the gateway into the fabric. It distributes inbound connections to the pool of command nodes. The network director also handles outbound traffic from nodes that have no direct external connection to the outside network.

The resource manager is responsible for allocating processors to user applications. Users and administrators interact with the resource manager to determine system status and to run large parallel applications on application nodes. I/O nodes provide file system access to command and application nodes. There are two types of I/O nodes. File system I/O nodes support one or more mounted file systems, either with directly connected disks or via a connection to a Storage Area Network (SAN). Network I/O nodes provide access to global NFS-mounted file systems and can attach to various customer networks with different interfaces.

Collage Configurations

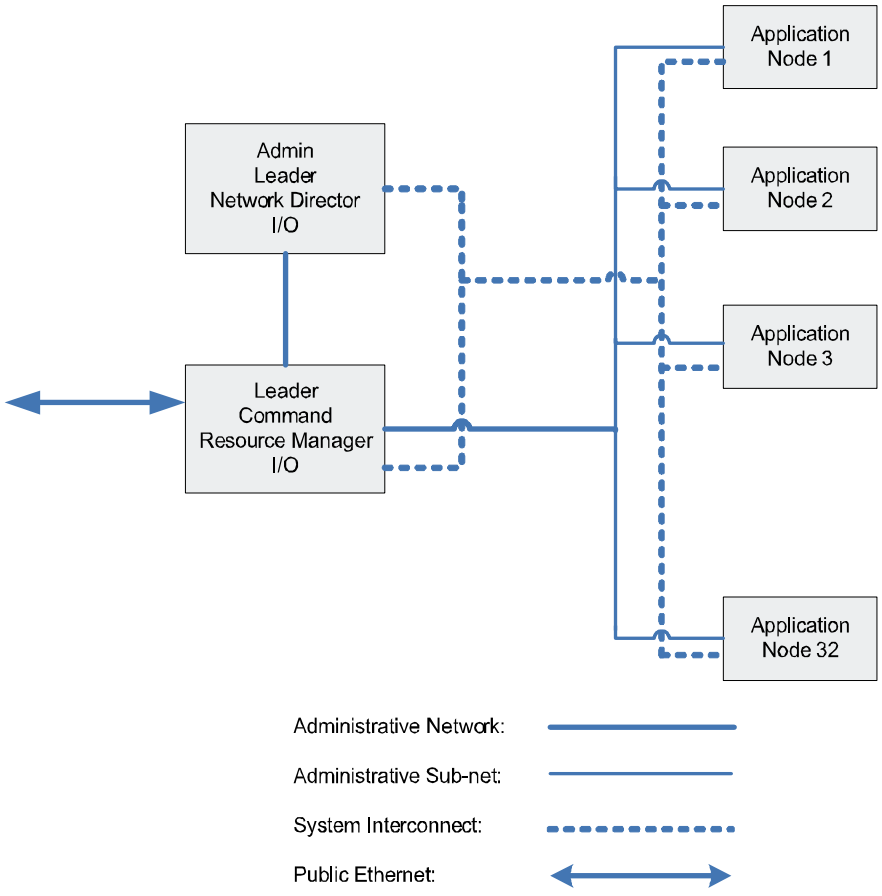
Collage utilizes two networks to manage cluster/grids. In Figure 1 and Figure 2, these networks are referred to as the *administrative network*, a hierarchy of sub-networks of service and application nodes, and the *system interconnect network*, which uses Gigabit Ethernet interconnect for inter-nodal communication among application nodes (for message passing and other functions).

The administrative network and the administrative sub-networks associated with groups of applications nodes comprise an Ethernet-based, private network. They provide connections from the admin node to all other application and service nodes, and in some configurations, to system components such as the system interconnect switch and power controllers. The administrative network uses NFS as a transport for distributing the shared root file system to all nodes.

Tel. (978) 263-3399
(978) 263-0033
hrg@hrgresearch.com

In Figure 1 (a small fabric of 32 nodes) and Figure 2 (a medium sized fabric of 64 nodes), the groups of 32 application nodes are members of administrative sub-nets that include a node with leader/command/I/O roles at the top. These leader/command/I/O nodes are connected via the administrative network to the admin node and to the resource manager. The resulting hierarchy, for each cluster, consists of three levels with the admin node at the top, the nodes with leader/command/I/O roles in the middle, and the applications nodes at the bottom.

Figure 1: A Small Collage Cluster



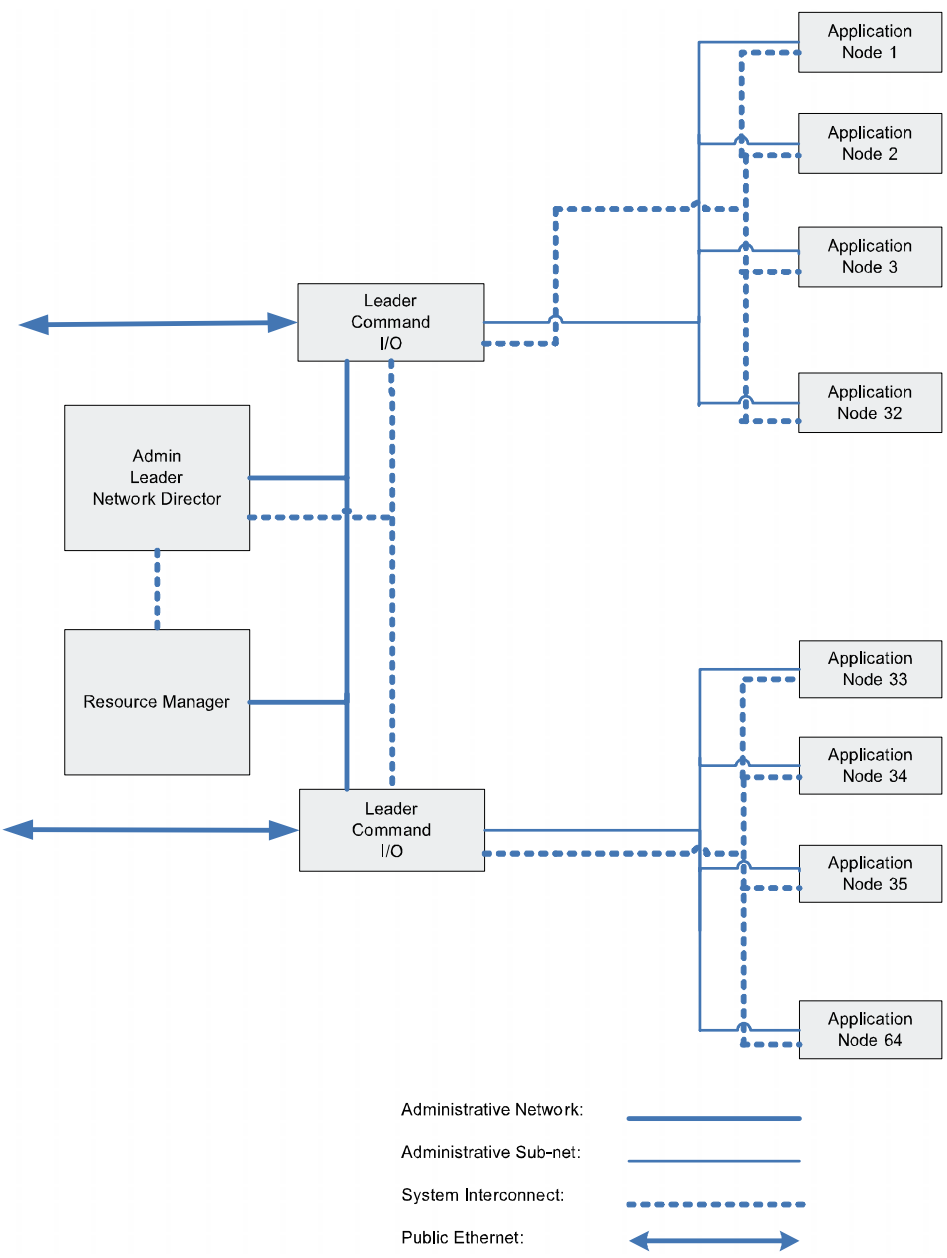
Source: Cassatt and HRG, June 2004

The system interconnect network provides transport for user application communication between the service and application nodes. Figure 1 and Figure 2 show all nodes connected via the system interconnect network. Applications communicate across this network using message-passing protocols such as MPI. The interconnect network is used for process management surrounding events such as launching applications and exiting applications. User I/O requests for files are also communicated across the interconnect network.

For the small cluster in Figure 1, the resource manager role is contained in the node with the leader/command/I/O roles. But for larger clusters, such as the one in Figure 2, the resource manager role is dedicated to a node and each cluster group has its own leader/command/I/O node. The resource manager in Figure 2 has more work to do with respect to allocating processors to applications than the resource manager in the smaller cluster — the reason for a node being dedicated to the resource manager role.

Tel. (978) 263-3399
(978) 263-0033
hrg@hrgresearch.com

Figure 2: A Medium Size Collage Cluster



Source: Cassatt and HRG, June 2004

System Resiliency

System resiliency is an important aspect of Collage, and it is expected to improve with future releases. Currently, the application, command, leader, and resource manager nodes can be configured to failover to a backup node without causing a system failure.

System Administration

A system administrator views Collage as an environment that can be used to manage various aspects of a cluster. Standard operating system administration commands are used to manage the rest of the system. Using Collage, a

Tel. (978) 263-3399
(978) 263-0033
hrg@hrgresearch.com

system administrator can control the number and types of nodes, install software (and software patches), monitor the health of nodes, support job queuing services, activate provisioning, and monitor audit trails of changes to nodes. By reducing the complexity that comes with increases in scale, Collage simplifies the IT lifecycle management process of patches, configuration control, and allocation.

Unified System Image

Collage provides the user (as well as system administrator) with a single unified system view, a single login, an integrated development environment, integrated job submission facility (such as Platform's LSF 5.0), and a single file system name space via the Unified Parallel File Service (UPFS). Collage also supports MPI for parallel applications, but other protocols can be supported for the development of applications.

UPFS is a high speed, scalable I/O capability that provides high performance I/O between nodes. It is capable of scaling to the bandwidth limits of interconnects such as Ethernet, Giga Ethernet, and is adaptable to emerging interconnects such as InfiniBand and PCI Express. Its global file system name space is a scalable solution that allows file systems from selected nodes to be available to the other nodes in a Collage cluster/grid. As a result, multiple file systems are treated as a single global parallel file system. The use of the single file system concept enables an entire cluster/grid to be viewed as a single unified system capable of supporting single user login, single file system name space, and single integrated view of resources.

Conclusions

Cassatt's Collage run-time fabric is an enterprise class application execution environment. It is suitable for High Performance Business Computing applications such as data mining, data warehousing and analytics and High Performance Technical Computing applications. Over the next couple of years, Collage will evolve to an application execution environment that supports transaction-based applications and SOA-based applications. It is HRG's perspective that Cassatt's vision and product plans will help enterprises move toward the goal of realizing on-demand computing.

Because the underlying architecture for Cassatt Collage is based on sophisticated role-based computing, the run-time fabric can grow without the bottlenecks that occur in most other cluster/grid environments. The Cassatt Collage run-time fabric improves price/performance by enabling IT organizations to replace large, expensive proprietary systems with a flexible application execution environment built on commodity components. Cassatt Collage effectively addresses one of the main challenges of IT managers — determining how to reduce cost/complexity while retaining the flexibility required to meet business needs.

Tel. (978) 263-3399
(978) 263-0033
hrg@hrgresearch.com

Harvard Research Group is an information technology market research and consulting company. The company provides highly focused market research, consulting services, and business modeling tools to vendors and users of computer hardware, software, and services. For more information contact Harvard Research Group as follows:

Harvard Research Group™
P.O. Box 297
Harvard, MA 01451 USA
Tel. (978) 263-3399
Fax (978) 263-0033

E-mail: hrg@hrgresearch.com
http://www.hrgresearch.com