

Under the Atlantic Ocean We Go, Asynchronously: Fujitsu Siemens Computers and EMC Combine Forces to Do the Job

Harvey Hindin, Vice President and Lead Analyst, Business Continuity ♦ March 1, 2004

Are organizations that need long-distance disaster failover and recovery as part of their business continuity planning looking for a demonstration showing that automated global application failover and recovery across the Atlantic is feasible with generally available equipment? It seemed like a good idea to Fujitsu Siemens Computers and Fujitsu, which are aggressively striving to increase their worldwide SPARC-based, Solaris-compatible computer sales and market share.

In a marriage arranged to perform just such a feat and show major-league capabilities, Fujitsu Siemens Computers (FSC) and EMC combined their respective computer/cluster and disk storage array-based asynchronous data communications capabilities. This combination proved that the Sales and Distribution module of the SAP R/3 application could recover from a simulated disaster in Hopkinton, Massachusetts in the U.S., and restart in Cork, Ireland in less than five minutes.¹

Architecture Overview

Making the under-the-Atlantic-trip required a PRIMEPOWER 400 server running Solaris 8 and EMC's Symmetrix Remote Data Facility/Asynchronous (SRDF/A) software running on Symmetrix DMX800 storage.² This Technology Trends provides an overview of the joint FSC-EMC achievement. Further details are also available.³

Table 1 provides a top view of the components of the architecture set up by the two firms (excluding networking and switching). Figure 1 (page 3) shows the simplified connection topology for the PRIMEPOWER 400 servers. Figure 2 (page 3) shows the simplified Symmetrix DMX800 SRDF/A connections.

Table 1: Key Architecture Components

Function	Implementation
<i>Node</i>	PRIMEPOWER 400 with Solaris 8, Emulex HBA
<i>Cluster</i>	PRIMECLUSTER with mySAP Wizard, SRDF/A Wizard
<i>Enterprise Application</i>	SAP R/3 IDES V4.6C on Oracle V8.17
<i>Remote Data Replication</i>	SRDF/A
<i>Storage</i>	Symmetrix DMX800
<i>Communications</i>	EMC native GigE MPCD; commercial routers, switches
<i>Load Driver</i>	Windows 2000 running SAP Sales and Distribution benchmarking; PRIMECLUSTER Administrative GUI; EMC management software

D.H. Brown Associates, Inc.

www.dhbrown.com

Our research program in Business Continuity makes this Technology Trends available to all our subscribers. Those interested in this program should contact cust_service@dhbrown.com.

Under the Atlantic Ocean We Go, Asynchronously: Fujitsu Siemens Computers and EMC Combine Forces to Do the Job

March 1, 2004

D. H. Brown Associates, Inc. (DHBA) Opinion

The fact that disaster recovery is on the minds of those IT professionals worldwide who are concerned with business continuity is not in doubt given today's world of cyber threats, bio threats, power outages, natural and manmade disasters, and a multitude of other ills that could strike without warning. Just what options can, or should be, investigated, invested in, or implemented to address these disasters is a subject of much interest in the IT community. While some organizations maintain disaster recovery infrastructure options for their Tier 1 (most business-critical) applications, many do not or are only in the planning stages (depending on the industry sector).

The SAP R/3 application is typically a Tier 1 application and represents a candidate for operation in a disaster recovery scenario. The significant element in this scenario – besides the quality of the demonstration and both firms' efforts to make it as real world as possible – can be found in the use of generally available equipment that can be purchased today. Those IT managers who have put together their own "specials" to accomplish similar feats will particularly appreciate such general availability. Finally, no one can say that Hopkinton to Cork is not a long distance.

DHBA believes that such "global recovery" will become increasingly important in the coming years to counter the threats businesses face. DHBA further believes that the trend of using generally available, lower-cost components in "standard" solutions should continue and provide affordable solutions in the future for mid-market enterprises as well.

The integration of replication with clustering technology enabling organizations to automate application availability for production and disaster recovery provides a powerful capability. The history of these efforts offers interesting insight into the current situation. Previously, these efforts typically allowed less than a couple of hundred kilometers between the primary and secondary locations. The limitation derived from the fact that as distance increases between primary and secondary locations, the propagation delay caused by the finite speed of light negatively affects local application response times. Moreover, clustering technologies are often limited by the distance limitations imposed by the cluster's underlying synchronous data replication and its software overhead. Finally, data consistency may pose a problem with certain asynchronous data replication methods.

The discussion of the technology behind what these firms did and how they did it rests on the information presented in Figures 1 and 2 (next page). Sidebar 1 (page 4) explains the network connections behind Figures 1 and 2 at a top level. The actual test data and the test procedures furnished in this document reveal the disaster recovery capabilities of this architecture and demonstrate the success of the joint FSC-EMC endeavor.

Under the Atlantic Ocean We Go, Asynchronously: Fujitsu Siemens Computers and EMC Combine Forces to Do the Job

March 1, 2004

Figure 1: PRIMEPOWER 400 Connection Topology (Simplified)



Figure 2: EMC SRDF/A Connection Topology (Simplified)



A Careful Sequence

In the first step of the demonstration, a failure was initiated at the Hopkinton site by having a member of the audience pull the power cable on the PRIMEPOWER server and the communication links from the Symmetrix DMX to simulate a site failure. The Cork PRIMECLUSTER node recognized the failure and initiated failover. This PRIMECLUSTER process involved the Cork Symmetrix DMX800 being designated as the new primary storage subsystem and the Hopkinton SAP application marked as down.

Step two covered the transition from Hopkinton to Cork. To do this, PRIMECLUSTER software in Cork confirmed the primary site failure and identified all the application's linked services and resource components. Then, the Cork SRDF/A activated its volumes. In addition, the Cork SAP instance was validated.

Step three dealt with the recovery at the Cork site. Cork site SAP services came online by means of PRIMECLUSTER software at Cork. After the Cork system accepted SAP user logins, full application production rolled into place.

Only four minutes and 35 seconds elapsed for production restoration. This was well within the test's recovery time objective⁴ (RTO) of ten minutes.

Under the Atlantic Ocean We Go, Asynchronously: Fujitsu Siemens Computers and EMC Combine Forces to Do the Job

March 1, 2004

Asynchronous Data Replication with SRDF/A

One of the key issues in obtaining an RTO of four minutes and 35 seconds is Cork's ability to maintain an almost-current copy of the SAP R/3 data running at Hopkinton so that the newly validated SAP instance at Cork can continue the Hopkinton work when the failover kicks in. EMC's SRDF/A technique for transmitting data asynchronously over long distances does the trick.⁵ How does SRDF/A do its job without requiring excessive bandwidth and the associated high costs? Moreover, how is this done with negligible impact on the running SAP R/3 application at Hopkinton?

SRDF/A provides remote mirroring (asynchronous replication) through a four-step procedure that is repeated constantly. Figure 3 (page 5) provides a top-level overview of the procedure, which is summarized here.

1. In Step One, the Capture Step, the SAP R/3 application-writes to the I/O at the active (source) site are collected (captured).
2. In Step Two, the Transmit Step, the Captured application-writes (a so-called "Delta Set" explained later) are sent to the disaster recovery (target) site.
3. In Step Three, the Receive Step, the Delta Set-writes sent in Step Two are received by the disaster recovery (target) site.
4. In Step Four, the Apply Step, the Receive Step delta-writes (after they have been received in their entirety) are applied to the disaster recovery (target) disks.

This four-step process is repeated as long as the primary site is functioning. Each time the cycle completes, a consistent, restartable data copy becomes available at the remote site.

The unique feature is the "delta writes" concept, which provides the benefit of locality of reference. Locality of reference takes advantage of an application's tendency to rewrite data to the same location multiple times. SRDF/A's Delta Set architecture enables data to be

Sidebar 1: Establishing the Network Connections

In preparing for a disaster failover, using the same equipment at both the primary and secondary sites is a good practice wherever possible. FCS and EMC followed this practice in this demonstration as seen in Figures 1 and 2. This symmetrical configuration offered a variety of practical benefits. For example, the host part of all IP addresses for the nodes at Hopkinton and Cork were the same. Similar benefits accrued to the WAN. Since a full-class IP address resides on each WAN side, simple configurations ensured routing between the two networks in Hopkinton and Cork.

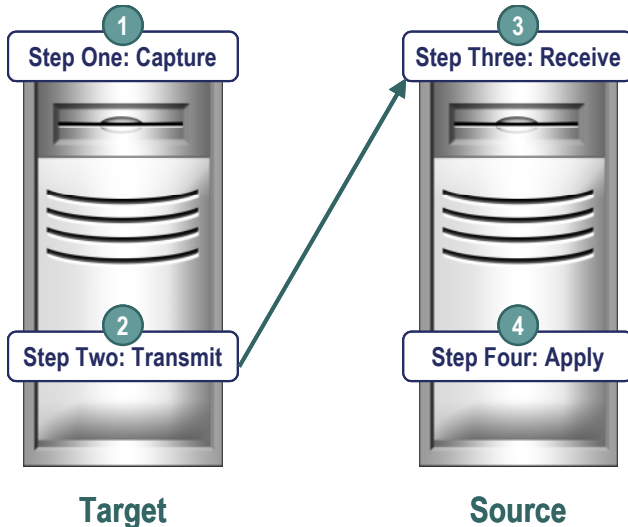
The Gigabit Ethernet switches used to provide WAN connectivity were employed by the PRIMEPOWER 400s and the two DMX800s. Multiple IP support on the switches was provided by VLANs. The switches worked with a router to furnish routing between the primary and remote sites.

SRDF/A was routed over Native IP. The needed Gigabit Ethernet connectivity was provided by two GigE Ethernet Multi-Protocol Channel Directors (MPCD) for Native IP in each DMX800 system. This eliminated the need for third-party protocol conversion devices. This approach also forms a best practice that may be reused at the Director level.

Under the Atlantic Ocean We Go, Asynchronously: Fujitsu Siemens Computers and EMC Combine Forces to Do the Job

March 1, 2004

Figure 3: SRDF/A Asynchronous Operation for Data Replication



set of writes is write-folded and transmitted to the target location. Normally, with asynchronous replication, every write created at the primary site is treated as a distinct write. It is time stamped, and sequenced before it is sent to the remote site. With SRDF/A, however, these writes remain in a cache (typically five to thirty seconds long depending on the specific configuration that the customer chooses to use based on the recovery point objectives and bandwidth requirements). This cache possesses special capabilities.

Writes to the cache that do not change during the Capture Delta Set are left alone. For writes that do change, only the new write is kept in the buffer. As a result, when the buffer contents are “promoted” and sent to the remote site, only the final update write set (in that five to thirty second period) travels over the WAN between the primary and secondary sites. (Note that the Capture Delta Set is not copied in cache. Instead, it is promoted in-place).

A new Capture Delta Set handles the next time period of writes and is itself promoted and sent to the remote site. Again, this happens in-place, in cache, to the Apply Delta Set. This approach can reduce the required WAN bandwidth considerably compared to the usual ordered asynchronous write procedure mentioned above.

The complex technology behind SRDF/A, and delta-writes in particular cannot be further elaborated on in this space. Potential issues with “out-of-order writes” at the secondary site may occur under certain circumstances. However, the SRDF/A software ensures remote

rewritten during a single Capture Delta Set step and only transmits the final set of writes to the target location.

In contrast, the traditional “ordered writes” approach requires every incoming write to be time stamped or sequenced and subsequently shipped to the target location. SRDF/A’s Delta Sets reduce the required bandwidth by cutting the amount of data that must be transmitted. It also is said to support “unlimited” distances with no performance impact.

Ordered or Delta Writes?

Delta Sets allow applications to rewrite to tracks numerous times during the Capture Delta Set. This design ensures that only the last

Under the Atlantic Ocean We Go, Asynchronously: Fujitsu Siemens Computers and EMC Combine Forces to Do the Job

March 1, 2004

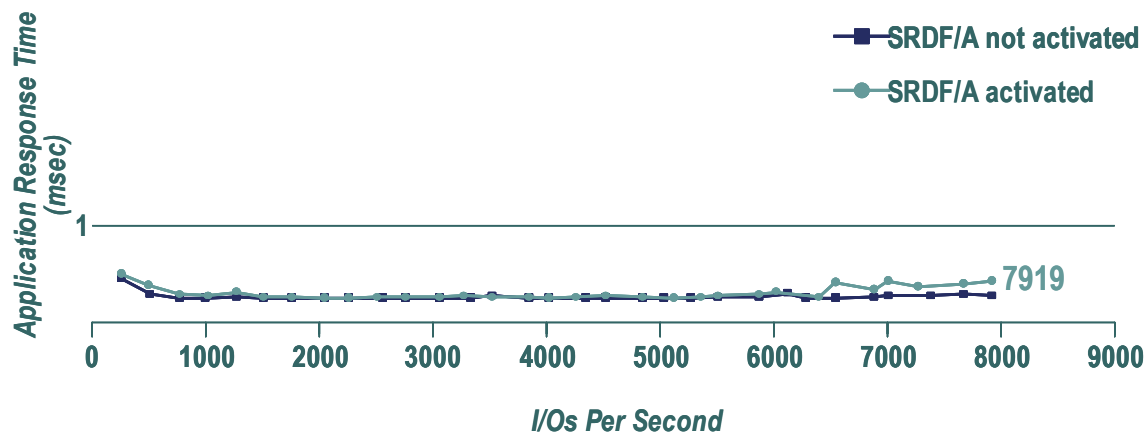
dependent write consistency by honoring the dependent write relationships embedded in the I/O application stream. More information is available from EMC⁶ or DHBA.

No Impact on the Primary Server

SRDF/A does its job without effect on the primary site server's application I/O response time. This protection of I/O response time underlies storage array-based replication – the storage does the work and the application does not suffer. Figure 4 (below) reveals proof of this thesis, which shows the SAP R/3 application I/O response time in milliseconds with the SRDF/A activated and with the SRDF/A turned off.

No difference exists for up to 6500 I/Os per second and there is a very small (unimportant) difference from 6500 to 8000 I/Os per second. The test was done for 100% sequential writes in 4K blocks.

Figure 4: Application I/O Response Time with and without SRDF/A



Test and Monitor the System

SAP R/3 is transaction-based and requires a full relational database.⁷ All its objects including users, printers, access rights, sales orders, invoices, etc. are stored in this database. Hence, SAP R/3 provides a good test of failover capabilities. The SAP R/3 Sales and Distribution (SD) subsystem was used for the cross-Atlantic demonstration. SAP provided a tool to allow an interactive load to be put on the cluster.⁸

A Windows 2000 Driver system, known as SD Benchmark, simulates interactive users accessing the SAP R/3 SD component. The tool put a load on the SAP R/3 system before, during, and after the failover. The success of this procedure demonstrated that the primary node first processed the load and, after failure, the secondary node processed the load as

Under the Atlantic Ocean We Go, Asynchronously: Fujitsu Siemens Computers and EMC Combine Forces to Do the Job

March 1, 2004

Sidebar 2: Wizard Wizardry Revealed

The philosophy behind Wizards is to reduce high availability or disaster recovery configuration times to days or even hours. This is done through the use of an almost turnkey software solution that only needs the insertion of application- or system-specific parameters. In many cases, the predefined menus and the reactions in the Wizard decision tables provide standardized solutions and minimize, or even prevent, implementation errors.

For example, the SRDF/A Wizard (jointly tested and certified by FSC and EMC) ensures that the communication links and Symmetrix device groups in a failover scenario function properly. One part of this chore is to write-enable the write-protected device groups prior to a failover, and resynchronize the original write-enabled device groups prior to a failback. The Wizard also prevents business processes from being switched over in a failure context after communication links have been broken, thus preventing the continuation of processes based on old data.

The SRDF/A Wizard in PRIMECLUSTER software implements failure detection and recovery schemes for SRDF/A environments. These schemes detect and automatically handle host, Fibre Channel interconnect, communication link, and storage device failure scenarios. They also intelligently minimize the number of switchovers. This capability proves key to minimizing overall business process disruption should a failover occur.

The SRDF/A Wizard supports a broad variety of storage interconnects and SAN topologies (non-cross-connected, single-cross-connected, and dual-cross-connected) with different multipath-drivers to balance specific environment requirements. This support of cross-connected configurations in local cluster environments helps avoid unnecessary failover. Such a capability is essential to maintaining business continuity – an issue that is today on everyone's mind.

FSC's PRIMECLUSTER contains Wizards for many applications including SAP R/3. The FSC mySAP Wizard enables the SAP R/3 application to be monitored so that it may enjoy Wizard-controlled fast failure detection and automatic recovery when required.

well. Furthermore, this approach ensured a realistic traffic load on the SAN to the primary DMX800 storage that generated traffic on the WAN (over SRDF/A) to the remote DMX800.

The Driver does even more. For example, it monitors the PRIMECLUSTER software status using the PRIMECLUSTER Administration GUI (through a web browser). The Driver is also connected to the SAN so that with the use of the EMC Symmetrix Administration software, it manages the DMX800s.

PRIMEPOWER Nodes and PRIMECLUSTER Activities

As mentioned earlier, it is PRIMECLUSTER on the PRIMEPOWER 400 server node in Cork that recognizes the failure and initiates the failover. The nodes in Hopkinton and Cork and their PRIMECLUSTER software, SRDF/A agent, and SAP R/3 agent play a critical role in the complete failover and recovery process. Also (noted earlier), this process requires

Under the Atlantic Ocean We Go, Asynchronously: Fujitsu Siemens Computers and EMC Combine Forces to Do the Job

March 1, 2004

the Cork Symmetrix DMX800 to be designated as the new primary storage system and the Hopkinton SAP application to be marked as down.

PRIMECLUSTER software in Cork then confirms the primary site failure and identifies the application's linked services and resource components. After this, the Cork SRDF/A replicated-volumes are activated and the Cork SAP R/3 instance is validated. During recovery at Cork, SAP services come online by means of PRIMECLUSTER software at Cork.

The PRIMECLUSTER SRDF/A Wizard capability stands out as a key component of the PRIMECLUSTER clustering software in this demonstration. PRIMECLUSTER Wizards are really GUI-based interactive tool sets that allow the creation and maintenance of disaster recovery configurations. *Sidebar 2: Wizard Wizardry Revealed* (previous page) provides further details.

PRIMECLUSTER Wizards

Clusters do not failover without a lot of software coming into play to make the transition smooth and to ensure that all the components that need to failover do so in a transparent manner. For PRIMECLUSTER Wizards, the general philosophy is to understand an application as a collection of components with their many relationships. The objective is to offer sophisticated mechanisms to handle any failure with minimal impact on running applications. PRIMECLUSTER Wizards provide the needed software to accomplish this transparent failover.

PRIMECLUSTER Wizards are an FSC-designed software infrastructure to configure applications for availability and scalability. Any application configured with PRIMECLUSTER Wizards is, by design, monitored using PRIMECLUSTER software's runtime cluster services.

PRIMECLUSTER's runtime services for high availability use the wizard-generated configuration (e.g., site variables and application policies) to ensure that an application and its associated resources are working.

Should a malfunction occur, PRIMECLUSTER's high-availability service manages, recovers, and/or switches failing applications and/or associated system components to preserve application availability. It accomplishes this by use of configuration policies and variables stored for the cluster site in question. These configuration policies and variables are set up using the wizards on a per-site basis.⁹ The PRIMECLUSTER Wizards discussed above include both standard, default, and certified templates, scripts, and detectors for configuring applications and their supporting resources.

Under the Atlantic Ocean We Go, Asynchronously: Fujitsu Siemens Computers and EMC Combine Forces to Do the Job

March 1, 2004

The PRIMECLUSTER Wizard architecture includes a base configuration engine, interfaces, a base generic application template, base sub-application templates, scripts, and detectors. All of these elements address the commonly available resources used by applications such as SAP R/3 and more.¹⁰ (The interaction of the Wizards and the high-availability software together with the use of the engine and various templates, scripts, and detectors requires a separate discussion not provided in this paper.)

PRIMECLUSTER Wizards also include customized application and sub-application templates, scripts, and detectors structured and preconfigured with default options to achieve the “best” implementation of specific applications.¹¹ In the absence of an already developed specific Wizard, an application may be configured for availability using the PRIMECLUSTER generic application and base sub-application templates, scripts, and detectors.

Final Thoughts

This Technology Trends has provided a top-level view of the joint effort of FSC and EMC to develop and demonstrate an asynchronous failover system between Hopkinton in the U.S. and Cork in Ireland. This noteworthy achievement expands in significance in light of how it came about. Only generally available components were used from FSC, EMC, and the communications vendors whose products were used. All too often, such installations require “specials” or are fully custom architectures. It is also praiseworthy that FSC and EMC have released a great number of technical details about what they did, and how they did it.¹²

Setting up such a system proves expensive. For example, the list price of the involved hardware and software components weighs in above \$1.3 M. However, recent user-requirements work by DHBA indicates that those organizations concerned with business continuity and disaster recovery (especially if such is mandated by government regulation) are prepared for this kind of expense.

DHBA believes that this demonstration will go a long way to helping Fujitsu and FSC meet their stated goal of doubling sales in the next few years. FSC has long commanded exceptional cluster technology, but until recently it has marketed PRIMECLUSTER more in EMEA and Pacific Rim markets by comparison with the marketing efforts of its competitors. PRIMEPOWER servers are sold through Fujitsu Computer Systems in North America, Fujitsu Siemens Computers in EMEA, and Fujitsu Limited in Japan and Asia Pacific countries. EMC Symmetrix storage is sold worldwide as well.

With this demonstration,¹³ FSC and EMC have shown that they can mount a major-league effort in North America and Europe. For example, the four minute and 35 second RTO across the Atlantic for the SAP R/3 SD subsystem is outstanding. As a result, DHBA

Under the Atlantic Ocean We Go, Asynchronously: Fujitsu Siemens Computers and EMC Combine Forces to Do the Job

March 1, 2004

expects that IT infrastructure managers will develop a better understanding of what FSC can do for business continuity and disaster recovery.

- ¹ Such distances have been handled and are handled in other venues by different vendors and IT infrastructures using a variety of technologies. These situations are not discussed here.
- ² Although this hardware and software was used for the demonstration, the capability extends to other PRIMEPOWER servers, EMC storage, and Solaris versions.
- ³ See, for example, Technical White Paper, *Global Recovery Demonstration: SRDF/A and PRIMECLUSTER* (available from FSC or EMC). Also, see FSC White Paper titled *PRIMECLUSTER Configuration Services*. There is also a PowerPoint presentation titled *Global Recovery: New Options in Meeting the Challenge* (available from FSC or EMC). Moreover, a large number of White Papers concern the PRIMEPOWER computers and PRIMECLUSTER. These were authored by DHBA and are available from FSC. Finally, additional information is available from DHBA.
- ⁴ Commonly used as a metric in both disaster recovery and business continuity scenarios, the RTO is the time (seconds, minutes, or hours) that is acceptable to come back online (for both application and data). The recovery point objective (RPO) is also often used as well and is the time point before the disaster to which data must be restored (seconds, minutes, or hours). This is equivalent to the acceptable amount of data loss. The specification of RTO and RPO is often part of an IT organization's Business Impact Analysis (BIA) performed as part of preplanning for business continuity and/or disaster recovery.
- ⁵ SRDF/A also works with mainframes.
- ⁶ See, for example, www.emc.com/products/software/srdf_a/interstitial/srdf_a_interstitial.jsp.
- ⁷ In this case, the database is Oracle 8.17.
- ⁸ Both the Hopkinton and Cork nodes had one PRIMEPOWER 400 each.
- ⁹ FSC says that PRIMECLUSTER Wizards provide a standard, all-inclusive approach to configuring any application. It further says that third-party application system components can be integrated using FSC's PRIMECLUSTER Configuration Definition Language (CDL). CDL is an XML-based language used to create templates. CDL is used by template writers to define the "user dialog," the relationship between resources, and other information required to create high-availability and scalable configurations of the desired resource or group of resources.
- ¹⁰ The base package is sold as one software license at a price dependent on the server model.
- ¹¹ Separate, customized Wizards are sold as separate software licenses independent of the server model.
- ¹² This is especially important in these days when certain accomplishments cannot be deeply analyzed due to the lack of information.
- ¹³ There are others not discussed in this document. Further information is available from FSC or DHBA.

This document is copyrighted © by D.H. Brown Associates, Inc. (DHBA) and is protected by U.S. and international copyright laws and conventions. This document may not be copied, reproduced, stored in a retrieval system, transmitted in any form, posted on a public or private website or bulletin board, or sublicensed to a third party without the written consent of DHBA. No copyright may be obscured or removed from the paper. D.H. Brown Associates, Inc. and DHBA are trademarks of D.H. Brown Associates, Inc. All trademarks and registered marks of products and companies referred to in this paper are protected.

This document was developed on the basis of information and sources believed to be reliable. This document is to be used "as is." DHBA makes no guarantees or representations regarding, and shall have no liability for the accuracy of, data, subject matter, quality, or timeliness of the content.