Technology Challenges for the Global Real-Time Enterprise.

Werner Vogels
Dept. of Computer Science, Cornell University,
vogels@cs.cornell.edu

Introduction

If there is one business concept that will drive distributed systems technology to the hilt in the coming years it is that of the real-time enterprise. The push for zero-latency access to a complete up-to-date view of all the business processes, internally within a corporation, as well as to customers, will dominate the thinking of system architects for the years to come. Many of the technology components that need to become the building blocks for constructing and managing the real-time global information flow within large corporations do not exist yet. It is certain that issues such as security, robustness, manageability, combined with legacy system integration, and all enveloped with a scalability coating will be central in the distributed system tools needed. It will require us to develop new technologies, re-package old ones, forge new tools and practices and to place these in an architectural vision in which many of these incompatible components can play together.

What is the real-time enterprise?

There is a strong push in business development to be able to react faster to changes within the corporation as well as in the world immediately around it. All of this is based on the notion that to remain competitive the enterprise needs to have a real-time view of operations internally, of the partners in the supply chain, of its customer's interest and of the competition, all to enable a more proactive management.

We have seen that the enabling of real-time operations in the front-office has led to staggering problems in other parts of the operations. This is largely because most of the business process remains implemented as a batch process. Customers are able to inspect goods and order online, but are no longer able to change orders once they have been accepted. This may be acceptable for direct sales to the end user but for the services that Dell Computer wants to offer to large distributors such as Frye's, who order up to 10,000 PCs at the time, this is not acceptable. Distributors must be able to change or augment their orders while they are being built, and the factory floor operation needs to be able to adapt to this instantly.

The pinnacle of real-time distributed business operations has for a long time been trading-floor automation. A matter of seconds was often the difference between success and failure. One of the buzzwords used in this context is that of the zero-latency, as in general the latency associated with the batch process, which still drives large parts of the business operations, inhibits the enterprise to react swiftly to changes internally and externally. Not that the trading environments were perfect as even though the trades may actually placed in real-time, the completion of the trade is a business process that can easily take up to 4-5 days be-cause of its batch style processing.

It is not only internally that enterprise needs to significantly improve its response times to events in the business process. For example the European low-cost airline ValueJet adjusts its fare pricing in real-time by monitoring the pricing of its competitors, to ensure that it offers the lowest fare, but not by too much. Almost the strongest push for real-time business operation comes from supply chain management divisions. In the past 5 years automating large parts of the supply chain have enabled businesses to reduce inventories significantly, making them leaner and more profitable. However it has also left them more vulnerable to shortages and hick-ups in the supply chain and the lack of buffering forces them to respond to changes at their suppliers in real-time.

An area where various experiments are underway is that of real-time Customer Relationship Management. The traditional approach of calculating "stable clusters" in a batch process at night is considered to be too conservative and does not meet the goals of the enterprise that wants to support customer business operations by online mechanisms. Instant monitoring of customer behavior to find relations with small variants such as changing a web-site background color or product packaging is considered essential for future adaptive on-line customer tracking. Targeted instant advertising as the customer is shopping online is considered to be one of the major growth areas for businesses.

It is not only traditional business that is focused on the reducing latency between the different information processes. The US Military is driving an effort that should also enable it to respond more quickly to change, whether it is the location of spare parts or the availability of new intelligence information. The Joint Battlespace Infosphere (JBI) architecture is one of the efforts to enable the global real-time information systems [1].

Other key markets that are adopting the real-time enterprise as a future business model are for example the telecommunications industry which will use real-time techniques for fraud-detection, and the airline reservation industry where every-thing is still main-frame, transaction driven but planners need real-time views.

For the coming 5 years Gartner estimates that for the multinational enterprise between 30 and 60 percent of the total IT budget will need to be assigned for developing real-time capabilities – equivalent to 1.6 and 3.2 percent of the total enterprise revenue. For a $5 billion a year business this translates into $80 and $160 million a year [2].

The different faces of the Real-Time Enterprise

While I am focusing on the technical aspects that underpin the real-time enterprise one must understand that more than technology is needed to make the transformation. Many business processes will need to be changed or adopted, and ultimately the decision will be human actions who must also adapt to the new timescale of doing business.

The examples in the previous section show three major aspects to the evolution of a corporation into a real-time enterprise:

- *Internal monitoring and information collection*: Each aspect of the business needs to be enabled to provide instant and up to date information to stake-holders, which maybe automated information fusion engines. Feedback cycles needs to be shortened and the control process needs to become adaptive to high-level directives.

- *Acquiring and responding to external events*: The ability to be proactive in business is necessary to create a competitive edge. Essential is that information about competitors and partners is available in real-time

- *Opening up the real-time process to partners and customers*: Customers, partners, investors, regulators, etc. will all demand access to real-time information from the corporation, which may have a different face depending on the receiver.

The Technology Base

The real-time enterprise is by nature event driven. This leads us to believe that the event notification technology is the natural candidate to help us to construct a loosely coupled distributed architecture that is flexible and extensible enough to support the evolution of the enterprise into the real-time information world. Event based systems allow us to construct the various components independent of the new or original implementations, which gives us simple paths for software up-grades and the insertion of new information management components. Essential for the real-time enterprise is that we can dynamically add information fusion components that consume event flows and react to the events or managed state.

Legacy applications can simply be extended into generating events, as the avoidance of synchronous operations does not alter the process flow of the original application. Nor does the introduction of new or updated components alter the flows in the system.

Technology Challenges

Even though the scope and urgency of building the real-time enterprise are spectacular by itself, at first sight the technology side is not very challenging. We have been building event based systems for workflow management and publish/subscribe systems for real-time data routing for some time now, and it appears a matter of technology integration to just extend these to other parts of the enterprise.

The main and major challenge however is that of scale. The real-time enterprise needs to be active at a global scope using the internet or internet like networks to support its main connectivity. The heterogeneity in connectivity as well as participant capability will require extremely flexible transport and content management techniques than that are currently not available in any single technology solution.

A second component of the scalability challenge will be that of sheer massive-ness of the amount of events generated at each moment anywhere in the global space of the enterprise. It will be impossible to take the traditional information bus approach as used in generic pub/sub systems and route the events anywhere in the enterprise. The architecture will need to include mechanisms for using information fusion components to reduce the overall event flow and to allow the system to move information fusion components throughout the network to places optimal to sources and consumers.

A third problem area related to scale is that the components can not predict the exact nature of the events they will need to deal with. In a global system where new event sources can come online each moment, new events type can be introduced any time. Requiring strict synchronization and registration will introduce too much coupling to build such a massive scalable system on, as such event processors will need to be flexible in dealing with a variety of data sources. It is likely that one can draw somewhat from the experiences with content based subscription and routing techniques, but none of these have been really applied to large business cases and at such large scale it is likely that significant additional research needs to be performed.

The fourth main problem is that of the robustness of the overall system. If the overall success of a corporation becomes depended on the timeliness characteristics of the information architecture it becomes essential that the overall system operates within clear bounds. Failures and disturbances are unavoidable at any time in a global operation, but their effect on the overall operation should be limited. A few slow or uncooperative components should not drag the overall performance down, and the effects of the saturation of parts of the infrastructure should be limited to its immediate locality and should not roll over into other areas. We will need to sort our refuge to probabilistic techniques for guaranteeing the data delivery, using probabilistic redundancy to achieve the needed robustness of these critical components. At the same time we need to establish mechanisms for assigning a reliability metric to information, such that fusion engines and/or the users can base their actions taking a certain level of uncertainty into account.

Essential to the overall success will be the scalability of the event infrastructure. Even though we can try to build the overall system as loosely coupled as possible, some infrastructure components will need to have information about the overall system. The event routing infrastructure will need to be able to potentially route events from any producer to any consumer in the system, based on quite complex specifications. This cannot be done in a scalable manner by simply using traditional routing tables, especially since there may be many possible receivers for any single event. The solution here lies in the use of routing hierarchies were each virtual node in the hierarchy contains the aggregation of the routing information in its children tables. Each router contains the tables between itself and the virtual root such that any given event can be forwarded those local consumers it knows about or to other routers that represent consumers with an interest in this event. For a description of these mechanisms and the epidemic communications that makes the maintenance of these tables scalable see the paper on the Astrolabe by Robbert van Renesse [3].

Even though event driven systems seem to address the needs of the real-time enterprise quite closely, there are a number of non-real-time requirements that may impact the technology that will eventually be seen as the best foundation for a solution. One of the related requirements is access to historical data. This can be an application driven requirement, for example when a context needs to be given for current events. It is also very likely that several legal requirements will exist that will force extensive logging and storage of event data. It is an active research area to determine where this data should be stored and how it can be accessed. Whether only raw data needs to be stored or that the data-fusion engines should also store their composite information streams, given that it may be difficult to re-play the causal relationships between the event streams. It is expected that relevant solutions will come out of the collaboration between distributed systems and data-mining researchers.

Related Work

There is a wealth of research and products that have generated technologies that will be part of the real-time enterprise architectures; Internet scale event messaging, high-performance content based routing, distributed subscription indexing, dynamic data-fusion engines, overlay networks, federated security, distributed data-mining, etc.

It would not do justice to the good work that has produced these technologies to provide a comparison or review at this point in the paper. What sets the new research directions apart from what has already been achieved is the focus on the problems of scale. In the coming years it will become clear that some of the technologies cannot make the transition to a large-scale environment, while others will be able to adjust to heterogeneous demands of scale.

Our group is using its many years of experience in reliable and scalable systems to build infrastructure support for the global real-time enterprise in the context of the Joint Battlespace Infosphere. This project is a collaboration between distributed systems, database and security researchers at Cornell. Our specific focus in this project is the scalability and reliability of xml-document messaging over a web-services based infrastructure.

Summary

In this short paper I have tried to give a brief introduction into the scalability challenges the research and development community face to enable the global real-time enterprise of the future. Even though there is a strong push for corporations to transition their operations into real-time, the current technology is by no means ready to face the challenges of true global scalability.

References

[1] United States Air Force Scientific~Advisory Board. Report on building the joint battlespace infosphere, volume 1: Summary. Technical report, SAB-TR-99-02, December 1999.

[2] Drobrik, A., The challenges of the real-time enterprise. Technical report, Gartner Research Report AV-14-9268, November 2001.

[3] van Renesse, R, Birman K. and Vogels W. Astrolabe: A robust and scalable technology for distributed system monitoring, management, and data mining. ACM Transactions on Computer Systems, 2003. researchers.