Enabling Omni Channel and Customer 360 with Enterprise Data Grids

Many large organizations are trying to merge data that's scattered across their business so they can provide unified access to the information they need in a timely and actionable manner. This includes static reference data and real-time data streams that are managed and moved by a wide variety of technologies, and owned by different lines of business. Here are a few examples:

- Banks maintaining a continuously updated 360-degree view of clients across all products (e.g. savings, trading, credit, insurance) and channels (e.g. web, branch, call center) so they can ensure satisfaction and take advantage of up-sell opportunities.
- Telcos providing a 360-degree view of client accounts and activity across products, services and channels, and reflecting in real time the sale and activation of mobile phones, changes to service plans, and utilization of services so they can extend, monitor and regulate usage, increase ARPU and improve customer retention.
- Transportation/shipping companies tracking the movement of vehicles and personnel so they can monitor and optimize resource allocation and system efficiency.
- Airlines reacting to the status and movement of planes and personnel, and availability of gates, ground crew and food services so they can optimize resource application, reduce costs and improve service.

For the past decade, most enterprise IT systems have been based on service-oriented architecture (SOA) and high-scale databases. Such systems can handle moderate data volumes and deliver services in seconds, but are struggling to support the volume of data generated and consumed by B2C applications, B2B partners and mobile computing devices. New technologies such as in-memory databases, database change data capture software, big data storage, analytics and complex event processing enable the creation of enterprise data grids that can consolidate siloed information into a single image that's updated in real time and accessible by many applications with very low latency. This performance and scalability enable a better user experience and give firms the flexibility to deliver new services that would not have been possible using traditional architectures.

This paper describes the architecture and applications of enterprise data grids, with an emphasis on the challenge of efficiently distributing information to consumers and between data grid instances.

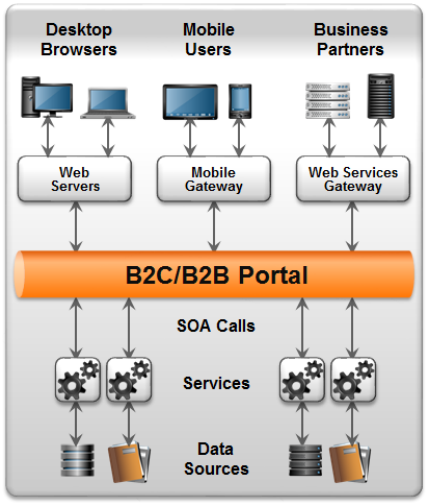
solace

http://www.solace.com

New Challenges for Legacy Architectures

External B2B and B2C interfaces have long been delivered using technologies such as Web Services, REST and web browser interfaces. There are many advantages and benefits of using these technologies to provide external interfaces and they serve this purpose well.

Many have tried to use the same technologies to pull information from internal systems as shown in the diagram on the right, because it lets the portal assemble a unified view for the external user by accessing well-defined interfaces without having to worry about the nature of those internal systems.



Portals usually need to call multiple services to satisfy each request. For example, a retail bank's web site needs to show clients the balances of accounts spread across multiple systems. That means each request must traverse the network, be converted by the SOA façade, fulfilled by the internal system and converted back by the SOA façade before the portal has what it needs to satisfy the user's request. Often these requests must be performed serially because the portal is single-threaded, or because information from one response must be used in the next request.

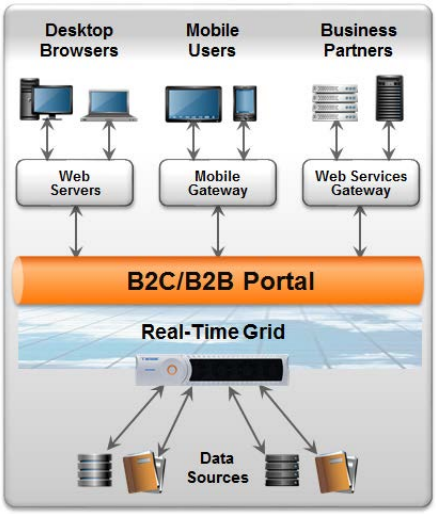
This "assemble on demand" approach introduces several problems:

- **Performance and Scalability:** The increasing number of users prompted by the success of mobile B2C applications, increased B2B partner interactions and the increased richness of information provided from these portals has resulted in external transaction rates going through the roof. SOA decouples the data access methods to internal systems, but those requests are still hitting applications that can't satisfy this increased rate without the help of query offload databases, caches or other complicated scaling solutions.
- **High Latency Response Times**. Acquiring data from many internal systems using SOA means multiple network hops through HTTP load balancers and other systems, introducing latency that's difficult to eliminate even if the internal systems respond quickly. This latency is even worse if the systems that must be accessed are located in a remote datacenter.
- **Lack of Real-Time Analytics**. Being inherently request/reply, SOAs do not support real-time – or even near real-time –analytics because data remains "at rest" until actively accessed by some consumer.

Solace can continuously update a real-time grid that lets a wide range of B2C and B2B interfaces present a unified view of information via the same web services they use today.

Scaling systems is expensive without enabling innovation. Failing to address these demands doesn't just cause poor user experience today, it limits the types of new systems that can be built for fear of impacting legacy systems.

The enterprise data grid architecture described in this paper accommodates today's requirements and provides a solid, flexible platform for the next generation of application development. A high-volume message bus such as Solace's can continuously update a real-time grid that lets a wide range of B2C and B2B interfaces present a unified view of up to date information and interactive services via the familiar web services technologies they use today.



Benefits of Enterprise Data Grids

The four main business benefits of an enterprise data grid are:

Unify View of Distributed Information

One of the most common applications of enterprise data grids is making data that lives in many siloed and geographically dispersed systems accessible as a unified 360-degree view of some aspect of operations, such as customer data and privileges. The data that goes in to such a unified view can be coming from systems of record via CDC, ETL, flat files, or a real-time JMS or AMQP feed of customer activity, usage records, etc. but is typically fed into the grid in real time as it changes in the systems of record.

Ensure Up-to-Date, Accurate Information

Some business information, such as customer address, changes infrequently and only needs to be reflected across systems within minutes of the change being made. Other changes, such as service activations and online payments, need to be updated everywhere as immediately as possible so they can be reflected on the next refresh of a web page, for example. The real complexity comes when you have a mixture of such changes, your infrastructure is distributed over the WAN, and you need to update different kinds of records with different latency SLAs.

Make Real-Time Data Accessible at Massive Scale

In addition to consolidating and synchronizing all of this globally distributed data, enterprise data grids can make it accessible to a massive number of applications. Low latency access to all of this information enables real-time services that don't degrade as application demands scale, and don't impact the performance of existing systems. This is a critical point not addressed by today's SOA implementations.

Perform Real-Time Analytics

Putting this enterprise information "in motion" allows analytics and complex event processing systems to identify meaningful events and patterns in streaming data – to add context to content. This enables business intelligence applications like fraud detection so firms can notify clients and proactively implement preventive measures. Another example is identifying dropped calls or slow website experiences so
they can reach out to customers to enhance or recover their experience.

Gartner on In-Memory Computing Technology

"In-memory computing is an emerging paradigm enabling user organizations to develop applications that run advanced queries on very large datasets or perform complex transactions at least one order of magnitude faster (and in a more scalable way) than when using conventional architectures."

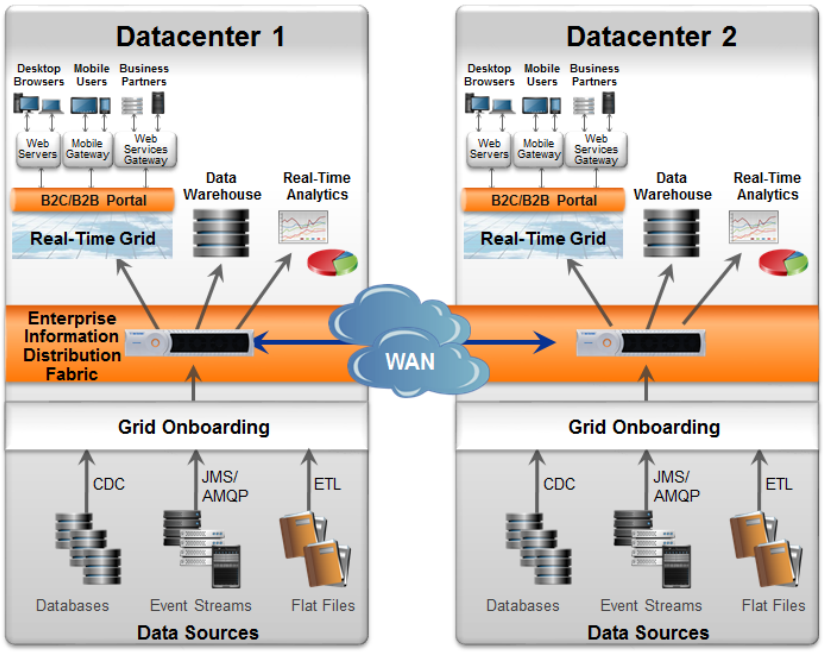
"The hype associated with big data is spilling over from MapReduce and Hadoop, which are technologies for big data at rest, to the world of high-performance message delivery products — which are technologies for big data in motion. High-performance message technology is part of the big data movement because of its ability to handle high volumes of data with high velocity processing and a large variety of message types."

Gartner *Hype Cycle for In-Memory Computing Technology, 2012,* July 2012

Next Generation Infrastructure for Enterprise Data Grids

To satisfy the business demands of modern IT systems, a consolidated view of enterprise information must be accessible to many applications at very high rates with very low latency. Increasingly, this information must also be made available in at least near-real-time to CEP and other Business Intelligence tools to provide near-real-time analytics. The architecture described in this section supports both requirements.

This diagram shows the basic architecture and components of next-generation enterprise-wide infrastructure, focusing on the functional requirements of the enterprise information distribution fabric.



Data Sources

At the bottom you see data sources—these can be combinations of master data repositories of existing siloed systems containing high value transactional data, reference data from 3rd parties and typical big data status or event streams generated in real-time.

Grid Onboarding

Various data source-specific adapters extract information from the data sources using technologies like CDC or ETL, or directly from real-time feeds and normalize it to a canonical model so it can be injected into the enterprise information distribution fabric, shown here being provided by Solace message routers.

Enterprise Information Distribution Fabric

The enterprise information distribution fabric (EIDF) is responsible for distributing data to many geographically dispersed systems including real-time analytics engines, big data technologies such as Hadoop and Cassandra, data warehouses, downstream applications and in-memory data grids for access by services and applications. The main functions required of the EIDF are:

- **Intelligent Routing:** The EIDF must use publish/subscribe distribution to selectively send to subscribing applications only the information they want. For example, a WiFi related app in a telco might get only the subset of information that applies to a customer's use of WiFi, while a fraud detection app might get a different set of real-time credit card transaction activity, but only for the clients it is monitoring.
- **Real-Time Delivery:** The EIDF must be able to satisfy the varying latency targets for SLAs of the different information objects while handling the presence of slow, offline and/or misbehaving consumers. This includes the ability to receive and absorb bursts from all of the grid onboarding agents, which could be used, for example, to enable a consolidated feed of call detail records or a user clickstream across multiple web applications.

- **Guaranteed Delivery:** The EIDF must provide lossless distribution of updates from the data sources to the real-time grid and any applications that require real-time streaming or asynchronous notification of changes. This delivery needs to meet relevant latency requirements, even during periods of peak volume. This allows the EIDF to be used for usage and other revenue generating transaction data for which the business cannot afford to operate on incorrect or stale information.
- **WAN Distribution:** The EIDF must be able to efficiently distribute updates across datacenters over long-distance WAN links. This must be done in such a way that all available bandwidth is effectively used and ideally eliminating the need for dedicated and expensive WAN optimization equipment.
- **Scalability:** The EIDF must be able to handle massive numbers of sources with ever-increasing volumes of information. This includes the ability to handle increasing numbers of subscribing applications which themselves must scale horizontally, where techniques such as sharding or striping of information while maintaining order of changes within a data object are required.
- **Security:** The EIDF must ensure authenticated and authorized access to the information and support encrypted transfer all without impeding the performance and scalability of the fabric.
- **Management & Monitoring:** The EIDF needs to let administrators monitor and manage the performance of application consumption and WAN delivery in real-time, and make changes on the fly. Increased visibility enables preemptive notification and action to avoid outages, and detailed debug data enables rapid root cause analysis.

Providing guaranteed distribution of updates from sources to the grid in real-time means the grid can be used for monitoring usage and other revenue generating transaction data for which the business can't operate on incorrect or stale information.

Consumers

The EIDF passes information to a variety of systems including:

- **Real-Time Grid:** The real time grid provides real-time application read access to a massive amount of consolidated information while staying in sync and accepting potentially high rates of data changes. This is typically implemented with NoSQL in-memory database technology to store data that must be accessed at high rates and with low latency, and then optionally a tier of high capacity and cheaper, but slower, storage for items that do not require the same level of read performance.
- **Real-Time Analytics:** One direct beneficiary of an enterprise information distribution fabric is analytics and CEP engines that watch real-time streams of data for events and patterns as they emerge so enterprises can better meet important SLAs by detecting and preventing fraud as it's attempted, notifying subscribers before reaching their plan limits, identifying cross-sell/upsell opportunities, and monitoring client activities across various points of access (web, partners, call center, etc.).
- **Data Warehousing:** The ability to capture data in motion and warehouse it for after-the-fact analysis of fault events, regulatory audit or to test new analytics in the future is now possible from the same information distribution fabric. Many such warehouse products are able to take feeds directly from a JMS API, making integration with the EIDF very simple.

By letting analytics and CEP engines watch real-time streams of data enterprises can detect fraud as it's attempted, notify subscribers *before* reaching their plan limits, identify upsell opportunities, and monitor client activities across various points of access.

Applications / Services

The applications supplied with information out of the real-time grid can take advantage of consolidated information that is always current, independent of the location of the source, and accessible at very high query rates and with very low latency. These applications can be for things like mobile service activations from partners, service changes, consumer or business web portals, dynamic notifications, or revenue generating applications. This includes applications that haven't been possible to date because it was too costly and/or difficult to access and update associated data. This is where the world of opportunity for business differentiation is opened up as the infrastructure barriers are broken down.

Solace offers unparalleled performance and scalability by handling both persistence and routing entirely in hardware.

Advantages of Solace's Approach

The enterprise information distribution fabric must receive massive amounts of real-time information and distribute it to many local and remote destinations while ensuring data integrity (in order, without loss). It can take enormous numbers of servers to do this in software, which is why Solace message routers are uniquely well suited to serving as the EIDF for next-generation enterprise data grids.

Massive Scale, Superior Performance

Solace offers unparalleled performance and scalability by implementing the data path of high-speed data persistence and distribution into hardware using patented techniques. As seen in the charts below, Solace offers much higher payload bandwidth and message rates than software in a fraction of the footprint. Depending on the scenario, Solace's solution is 10-50 times faster than the competition.

Over the WAN, two or more message routers can act as a distributed message fabric and provide a number of additional data grid performance enhancing functions such as:

- **Non-blocking Replication:** Solace stores messages then pipelines messages to the remote site (vs. synchronous request/reply) thus reducing the impact imposed by the high round trip time over the WAN.
- **TCP Optimizations:** Solace's technology includes optimizations such as large TCP receive windows and control over TCP slow start to maximize the throughput of individual TCP connections of high bandwidth links.
- **Streaming Compression:** Solace's ability to perform streaming payload compression in hardware provides on average an 80% compression ratio, saving bandwidth and reducing the latency of large message transfers, especially over low bandwidth links.
- **Striped TCP Connections:** By striping communications across multiple TCP/IP connections, Solace reduces the limitations of TCP/IP sliding windows over high latency networks.
- **Edge Cache:** Through pre-fetching and pre-emptive distribution of persistent data (as in edge computing or content delivery networks) enables Solace to replicate data across the WAN prior to system startup.

- **Duplicate Data Reduction:** Solace message routers only send updates once per link even when there are multiple receivers at the far end of the link. Remote message routers fan out messages to local recipient applications at the edge.
- **Bandwidth Throttling**: Solace tunnels all WAN traffic through a configurable set of TCP/IP ports. In conjunction with your existing IP routers and switches, these ports can be set to be rate limited so that they do not use 100% of the WAN bandwidth at the expense of other applications.

High Availability

Since enterprise data grids support important time-sensitive internal functions and customer-facing services, it's essential that they be absolutely reliable. Toward this end Solace message routers support high-availability with the following features:

- **Fault Tolerant Components:** Solace message routers are equipped with redundant fans, disks and power supplies to ensure continuous operation in the event of a component failure. This is combined with integrated monitoring and notification to management systems of hardware faults.
- **Fault Tolerant Connectivity**: Solace message routers provide fault tolerant connectivity to external components using Ethernet bonding, redundant SAN connections, and redundant connections between message routers in order to seamlessly provide service in the face of external equipment faults.
- **Integrated Fault Recovery Mechanisms:** Solace message routers have built-in mechanisms for 1+1 fault tolerance without external clustering software. For example, two message routers can be configured as a fault tolerant pair. The message routers automatically synchronize messages and monitor each other's health. If the primary fails the secondary automatically picks up where the active left off, commencing the delivery of messages in just a few seconds. This is all done with no risk of data loss or reordering, regardless of message load or how much persistent data is stored, which is not possible with software systems.

Solace message routers have built-in fault tolerance. Two message routers configured as a fault tolerant pair automatically synchronize messages so if the primary fails the secondary picks up in just

Rich and Simple Monitoring and Management

Solace message routers export granular status updates, statistics, rates and high water marks across all layers – something not possible with software implementations. For example, for a particular client connection, the administrator can see current and high water mark message queue depth, input/output message rates at 1 second and 60 second granularity, TCP round trip time, lost TCP packets, TCP re-transmits and more. You can list the publishers publishing the fastest, or the subscribers receiving at the highest rate. Similar information is available for each persistent queue. This information is always being gathered and does not impact the performance of the system due to the parallelization that's possible with a hardware solution.

This information is always accessible to management systems without impacting the performance of the message router. Administrators can see what is happening in the system even under peak load for troubleshooting and these metrics can be polled aggressively and put into a database or plotted in real-time without impact to any applications.

Solace message routers export granular status updates, statistics, rates and high water marks across all layers – something not possible with software implementations.

This rich status information is made available via a GUI called SolAdmin, command line interface or a RESTful XML over HTTP interface that custom management applications can use for provisioning or status gathering. Solace message routers can also be monitored using 3rd party monitoring products such as ITRS Geneos and Nastel. Monitoring of hardware events can be performed by custom or 3rd party or custom applications via SNMP or SYSLOG, and asynchronous events generated by the system, such as queue full warnings or queue overflows, can be received by applications via SYSLOG.

Proven Integration with 3rd Party Grid Components

Solace has successfully integrated many 3rd-party components of the enterprise data grid such as Oracle Golden Gate, Oracle Coherence and IBM Extreme Scale, Streambase, TIBCO BusinessWorks and all mainstream JEE containers for custom applications.

Shareable Infrastructure

Solace supports the creation of "Message VPNs" that give many applications and departments their own compartmentalized environment within a single Solace message router. When these Message VPNs are set up messages don't cross between environments unless you want them to. Administrators can fully configure the behavior of, and resources available to each virtual messaging partition. This can reduce TCO and accelerate time-to-market by enabling many applications to tap into the power of a single easily-managed device. Virtualization also allows several discrete development and test environments to be run on one message router.

Web Streaming

Solace message routers can stream data to Rich Internet Applications (RIA) using WebSocket in HTML5/JavaScript, Silverlight and Flash and mobile phones and tablets without requiring additional web streaming technology integrated with a different messaging product. This integrated capability can be used to provide mobile and web applications to external users and for internal dashboards and monitoring of the enterprise data grid itself.

Summary

Driven by increasing business demands and the availability of innovative new technologies, large organizations are turning to enterprise data grids to accelerate, optimize and scale their IT infrastructure. This new architecture is helping enterprises better compete in the global marketplace by improving the performance of application services and giving the business an agile architecture so they can quickly create new revenue generating services and monetize their information assets and customer interactions.

Central to any enterprise data grid architecture is the enterprise information distribution fabric, the layer responsible for efficiently moving, filtering and synchronizing information between data grid instances, analytics and data warehouses over the LAN and WAN in real time.

Solace message routers meet this requirement with higher performance, lower TCO and less complexity than any other solution—commercial, open source or home grown.

To learn more visit solace.com or call +1 613-271-1010.