Notes.net Iris Today

Home Download Iris Today Iris Cafe All About Domino Iris Sandbox About This Site

Russ Holden: Domino R5 Database Improvements

Interview by Betsy Kosheff

Level: All
Works with: Domino 5.0
Updated: 05/03/99

Inside this article:

Related links:
Domino R5 Technical Overview

Developer Spotlight

IBM's Almaden Research Center

Get the PDF:

RUSS.pdf (66Kb)

*[Editor's Note: To learn more about Domino R5 database improvements, check out the discussion with Russ in the **Developer Spotlight**.]*

Iris has been working for over two years to perfect the improved database that will debut with the new Domino R5 server. For team leader Russ Holden, it's all about the enterprise, making Domino even better at reliability, manageability, scalability, and performance, performance, performance.

What are the major areas of database improvement for the Domino R5 server?
Overall, we've accomplished some major improvements in database scalability and performance for the enterprise mail customer. There are also a lot of new features like transactional logging, online compaction, and high-performance online backup and restore, to name a few. And, customers will see significant improvements in view updating and rebuilding -- in the 3-5 times range for view updates and 3-10 times faster for view rebuilds.

What's driven your focus on mail database performance?
People want to know that they will get the bang for the buck if they add more memory or put another processor on, so they get back what they paid for. Previously, we hadn't been able to take advantage of multiprocessors very well, so we've worked on that a lot because customers are trying to consolidate servers and bring down the overall cost of managing their mail use.

Are there other new features for messaging users?
In addition to core enhancements to the data store, we've done a lot of work on performance of the router, especially as it uses the data store. Most of those enhancements come for free, but we've also added some options for the large enterprise. One new feature is the ability to have multiple mail.box databases, which reduces contention for heavily used mail servers.

How does it work?
The way Domino works today, we have one database called mail.box,

where all mail is deposited. This can become a bottleneck, especially if you're running a lot of users or if you want to compact the database. So, now we've allowed you to have any number of mail.box databases, so you can spread the load and the contention over as many as you wish. You can also take them offline individually. This has been one of the biggest ways we were able to improve scalability and really explode the benchmarks.

What does transactional logging bring to the enterprise user?
It's an industry-standard technique and really state-of-the-art for reliable data storage. A transactional log provides a sequential file that everybody writes to -- sequential writing on disk is much faster than writing in various places on disk.

When you perform a database operation, that operation is either completely successful or not successful, with no in between. Moreover, when you're done with that operation and you've received a response back, it needs to be guaranteed that the operation is recorded on disk, not just in memory. Transactional logging is what records the operation you performed. You could just choose to get the operation out to disk using the database, but that ends up being too expensive because there are a lot of things you need to update.

What other benefits do you get from transactional logging?
It enables you to perform a system restart without running DbFixup on your databases. This is a great performance win.

When the server crashes today, we run Fixup on all the databases that were changed since the time it crashed. In Domino R5, we use the logs to replay all the changes that didn't get into the databases, rather than running Fixup on each of them. The result is that restart is dramatically faster because you won't have the Fixup process running on big databases, which can be very slow. We just replay history through the log and bring everything up to date.

So, restarting the server is going to be faster?
Definitely -- in fact, restart of the server when logging is enabled should be orders of magnitude faster.

The other big benefit we get from logging is improvements in backup, because now we have a full set of application programming interfaces (APIs) for backing up databases. Domino R4.x does not have a backup and restore API, although there are various backup and restore tools available from different vendors. Rather than shipping a product that performs backups -- a lot of products already do that extremely well -- we're going to provide a set of industrial-strength APIs that allow vendors to hook in their products. This is going to give customers the high performance online backup and restore programs they need for Domino and Notes.

How will this change the process of performing backups?
Administrators will back up their full databases initially, then after that, they'll use the logs to record changes. So you might, for example, once a week or so record the full database and then you just need to use the log. It provides a very nice mechanism for ensuring that you haven't lost anything, without having to go and copy gigabytes of databases onto tape every night.

Are you going to work with specific systems management products?
We're working with about seven vendors who have existing Notes backup products. Now we're giving them the hooks they need to get inside the server, so the server will perform a lot better. Our ship date will coincide with the shipment of as many of these third-party products as possible. Each one of them will conform to our API.

Have you made changes to the database size limit?
Yes. The new database architecture includes a change in how we store pointers inside the database and this allows us to provide individual databases over one terabyte in the near future. For R5, we will test and certify databases up to at least 64GB, versus the R4.x limit of 4GB, and we'll increase that number rapidly in the future.

What's new about compaction in R5?
The purpose of compaction is to reclaim space on the database -- for example, for documents that have been deleted. The way Domino works, rather than making the file smaller when you delete things, it will wait and try to reuse that space later. Consequently, sometimes there's an accumulation of free space, so you need compaction to reclaim it. In Release 4.6, compaction required you to make a new database copy. If you had a 4GB database, you needed another 4GB of space on that same device. That was very slow. In R5, once you have your database in the new format, compaction is done in place. We do all the work of shifting things around within the boundaries of the file that's on disk. So you don't need all that extra space, plus it's dramatically faster.

However, be aware that when you do perform compaction the first time, you have to copy your databases; but once you do that, you get all the new functionality and performance for that investment.

Can I do my compaction while users are online?
Yes, you no longer need to have exclusive access to the database to perform compactions. Previously, no one else could be reading the database or updating it; but with Domino R5, users can continue their operations while the compaction process is underway. Administrators don't have to shut down or figure out how to get exclusive access to the database.

There's also a new On-Disk Structure (ODS)?
Yes, and it's going to net some big performance improvements for people who upgrade. At the same time, all old formats will still work in R5.

What was the thinking behind the new ODS?
We've been iteratively changing the format of the database to optimize I/O and CPUs for mail and other application workloads. The idea is to make it so that when you add mail, for example, we have to write fewer things to disk. Then when we do write to disk, we write in a smaller number of places on the disk and do much more sequential writing. What we get out of this is approximately half the I/O rate as in R4.x databases for the same set of operations. And, in some cases, it's actually much less than that. The new ODS is also what allows in-place compaction and transactional logging recovery to take place.

Are there other new database functions?
There are a number of new features that make certain functions optional, particularly those which may take up system resources.

For example, we now have the ability to create databases that don't maintain unread marks. It's important to have the capability to keep track of documents you haven't read, but for some databases, like the Domino Directory or log files, it's not something you typically need. So, now there's a switch you can use to turn it off, and that's giving us some additional performance improvements.

There's also the ability to turn off overwrite, if physical security is already adequate. The Release 4.x server allowed for all deleted data to be overwritten on disk, which has a performance cost. So now that can be optional.

How many users will you support on a server?
Most customers today have between 100 and 1,000 active users on a single server. But we do have many large sites that support far greater numbers -- in one case, 10,000 users on a single server. With R5, our current informal analysis indicates most customers will fall in the range of 500 and 5,000 users per server, although we expect to be able to handle far larger numbers than that, probably in the 10,000 range.

Of course, the real measures of scalability are not just the number of users per server, but the total infrastructure needs of a large organization. That means multiple directories, distributed administration, flexible authentication/encryption, and replication/routing technologies that remain bulletproof over any network topology limitations. These are equally important things that people should be looking at when they make comparisons.

What else did you improve in R5?
We've been working on and accumulating improvements to the data store for about 27 months now, so we've had a lot of time and we feel pretty confident the work we've done is going to put us out ahead of the competition in a number of areas. Overall, we have re-architected the internal structure of the data store, which has allowed us to achieve major enhancements in CPU and I/O utilization, as well as major improvements in database integrity.

What are some of the specific areas where it will be better?
We gain some of the improvements from the new data store, while others have been inherent features of the server for some time.

Specifically, our mail files are easier to handle as individual entities, which is particularly useful for administrators performing user moves and individual mail file recovery. Exchange requires you to restore entire mailstores in order to manually extract individual mail file data.

Also, our database was built from the ground up for self-describing semi-structured data, so we don't have some of Exchange's relational database limitations, such as a directory that is limited to 10 custom fields. Our database access control granularity and replication mechanisms are also far superior, as is clustering and partitioning. Exchange has only two-node failover, while we have 2-6 node failover, plus load balancing, which works across platforms, not just on NT.

[Editor's Note: Node failover is a bit of a misnomer. It is actually client failover. This means that when a client tries to access a database on a server and that server is unavailable, or has a higher load than it should, the client fails over to (or accesses) a server that has a replica of the same database on it. Domino clusters provide this mechanism, as well as mechanisms to keep the replicas of databases on all the servers in a cluster in tight synchronization (to within a few seconds), so that clients can't see differences between replicas.]

Finally, we've married the flexibility and power of the data store with the benefits of traditional database logging and recovery mechanisms. The logging and recovery work has been done by teaming with a group of top-flight developers and researchers at **IBM's Almaden Research Center** . This is the place where relational databases were invented and where major new database technologies for data warehousing, data mining, multimedia and Web-enabled applications are being perfected. So we are benefiting from IBM's experience with many real-world enterprise customers here.

Is there going to be a difference in the scalability of R5 servers on different platforms?
Now that we have largely eliminated bottlenecks in Domino itself, and scale dramatically better on multiprocessors, we are working to exploit those changes so that the power of some of the large-scale Unix systems, AS400s and OS390s, provides better value for the price and performance level than NT-based systems.

What should people know about upgrading databases to the new format?
The process of upgrading is the same as it's always been, which is that you compact databases over to the new format. We always allow customers to retain the old version of the database. You don't have to compact them, you can leave them in the old format and we'll always support it. But there are a lot of functional reasons why you're going to want to upgrade -- for better scalability, for the new transactional logging, and, of course, better performance.

BIOGRAPHY
Russ Holden is in the Database Project Group at Iris Associates. He has been working on the redesign of the Notes database system, broad server performance, and scalability improvements, since just after R4 shipped. There are currently about 12 people working on the database management system alone. Russ has been at Iris for almost four years. Prior to that he was at Digital for seven years working on Distributed Database Management Systems (RdbStar and DB Integrator projects), and at CCA (Computer Corporation of America) also working on a Distributed Database Management System (Adaplex). His hobbies are running, cycling, and racquetball. Outside of work, almost all of his time is taken up by his best buddy (usually!) - his four-year-old son, Brandon.

What do you think about this article?

Register Here!

Lotus Home | **IBM Home** | **Iris Home** | **Feedback**