

IBM

ISSUE 1, 2010 / VOLUME 15 NO. 1

data management

IBM.COM/DMMAGAZINE

KNOWLEDGE. PERFORMANCE. RESULTS.

1011011000 011100011 0010
1011011000

extreme

databases:

the biggest
and fastest

What you can learn from
life on the database edge

DON'T BLINK

IBM solidDB has
already completed
your transaction

CALL ME

How telecom connects
in real time, every time

STAY CONSISTENT

The inside story on
DB2 isolation levels



A Hercules Beetle can lift objects weighing over 300 times its own body weight.

AMAZING FEATS

**BUSINESS RUNS ON I.T.
I.T. RUNS ON BMC™**



A mainframe running BMC SQL Performance for DB2 can process 50% more without extra MIPS.

OF STRENGTH

» Watch the 2-minute overview of
BMC SQL Performance for DB2:

WWW.BMC.COM/STRENGTH



 **bmcsoftware**



data management

KNOWLEDGE. PERFORMANCE. RESULTS.

ISSUE 1, 2010 / VOLUME 15 NO. 1



Features

32 **solidDB** and the Secrets of Speed

How the IBM in-memory database
redefines high performance

35 **What Is DB2 pureScale?**

Going to extremes on scale
and availability for DB2

extreme 18 **databases:** the biggest and fastest

Learning from life on the database edge

2010

IIUG Informix Conference



April 25-28

Overland Park, KS (Kansas City) | Overland Park Marriott

April 25 – In-depth tutorials
April 26-28 – Conference sessions



Join Informix users, developers
and enthusiasts, along with IBM
engineers and executives



For more information and to register visit <http://www.iiug.org/conf>
Use registration code **MAGDISC100** and save \$100 through April 21, 2010

Departments

6 Editor's Note

By Cameron Crotty

10 NewsBytes

14 IIUG User View

By Stuart Litel

15 IDUG User View

By David Beulke

16 Data Manager

Exploring the Extremes of Database Growth

By Merv Adrian

40 Data Architect

Making Humongous Doable

By Robert Catterall

44 Distributed DBA

Changes to the Cursor Stability Isolation Level: Part 1

By Roger E. Sanders

46 Informix DBA

Fastest Informix DBA Contest II: How Did They Do It?

By Lester Knutsen

48 Smarter is...

Understanding the Universe

By Chris Young

Ad Index

BMC Software **C2-1**
www.bmc.com

DBI **C3**
www.DBISoftware.com

IBM **8-9**
www.ibm.com

Intel **38-39, C4**
www.intel.com

International DB2 Users Group **7**
www.idug.org

International Informix Users Group **3**
www.iiug.org

ITGAIN **43**
www.itgain.de

Quest Software **37**
www.quest.com

Relational Architects International **23**
www.relarc.com

Responsive Systems **5**
www.responsivesystems.com

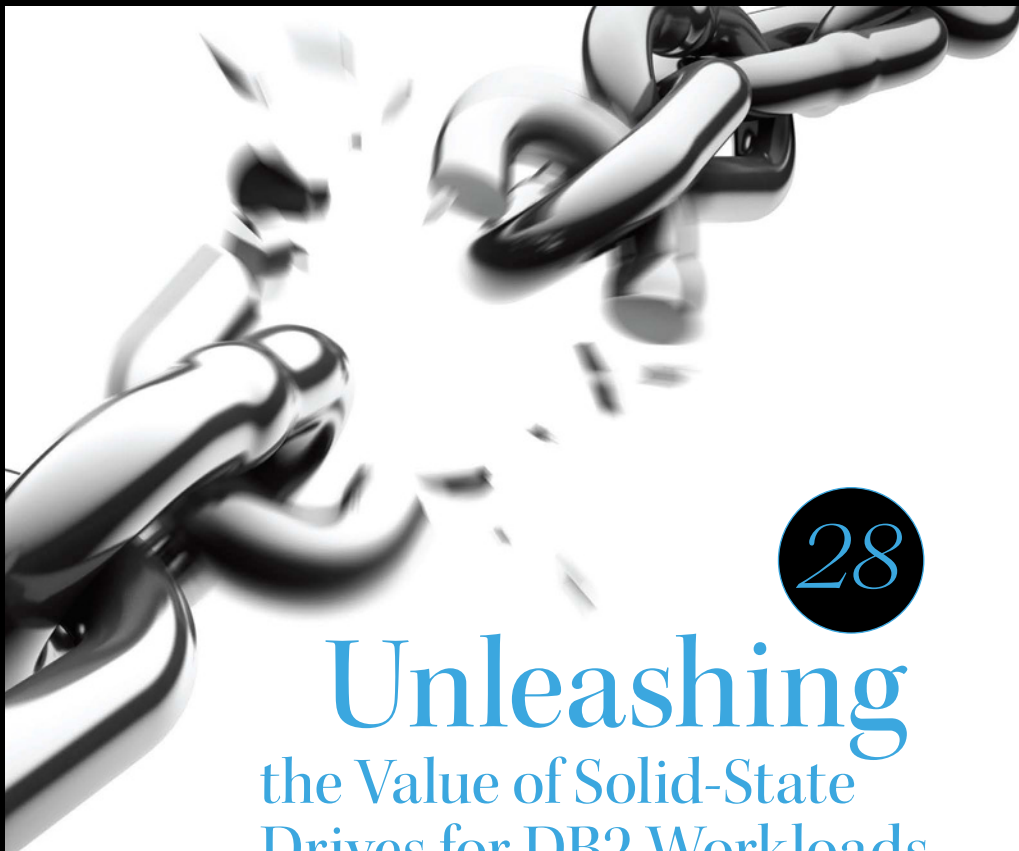
Vision Solutions **42**
www.visionsolutions.com

24

Industry Focus:
Telecommunications

Call Me

What's the most important part of a telecommunications business? Data.



28

Unleashing the Value of Solid-State Drives for DB2 Workloads

A little solid-state speed can go a long way—here's how to make the most of low-latency storage

Optimize Both Performance & DB2 Memory Utilization With

Buffer Pool Tool® For DB2

Reduce Your Processing Costs, Save \$\$ Now!

PRODUCT FUNCTIONS	BUFFER POOL TOOL	BMC POOL ADVISOR	IBM POOL ANALYZER	UBS HAINER BPA4DB2
Collects trace data for the critical performance period	✓		✓	✓
Easily processes up to 16GB of trace data	✓			
Huge performance trace must be down-loaded to the Workstation			✓	✓
Simulates/predicts pool & object I/O performance	✓			
Simulates/predicts object partition usage & I/O rate	✓			
Simulates I/O for moving objects to other (& new) pools	✓			
Simulates/predicts Group Buffer Pool performance	✓			
Proven I/O rate prediction accuracy, ask our clients	✓			
Provides a <i>proven methodology</i> for pool tuning	✓			
Proven low CPU overhead for data collections	✓			
Continual CPU overhead on your processor		✓	✓	✓
Reports memory usage & paging of DB2 & entire LPAR	✓			
Shows you how to reduce transaction elapsed times	✓			
Windows based graphic performance analysis	✓		✓	✓
Performance Tuning Wizard	✓			
Reports CPU costs by object for I/O and Scan activity	✓			
Easily installs in less than half an hour	✓			
Client references that will discuss their savings	✓			

**RESPONSIVE
SYSTEMS**

281 Hwy 79

Tel: 732 972.1261

Web: www.responsivesystems.com

Morganville, NJ 07751

Fax: 732 972.9416

IBM data management

KNOWLEDGE. PERFORMANCE. RESULTS.

EDITOR

Cameron Crotty
editor@tdagroup.com

MANAGING EDITOR

Stephanie S. McLoughlin

ART DIRECTOR

Iva Frank

DESIGNERS

David Chan, Lalaine Gagni, Margie Preston

CONTRIBUTING WRITERS

Merv Adrian, Eric Ahrendt, Bonnie Baker, David Beulke, Scott Bisang, Robert Catterall, Chris Eaton, Tam Harbert, Sally Hartnell, Sunil Kamath, Lester Knutsen, Stuart Litel, Roger E. Sanders, Antoni Wolski, Chris Young, Paul C. Zikopoulos

EDITORIAL BOARD OF DIRECTORS

Rick Myllesbeck (Chair), Scott Bisang, Debra J. Black, Cathy Elliott, Jeff Jones, Kimberly C. Madia, Nancy Miller, Jenn Reese, Bob Sawyer, Lindsay M. Scariati, Kathryn Zeidenstein

AD SALES EXECUTIVE

Randy Byers
advertise@tdagroup.com

AD COORDINATOR

Katherine Hartlove

SPECIAL THANKS TO

Lea Anne Bantsari

SUBSCRIPTION SERVICES AND REPRINTS

To subscribe to the print or digital version of *IBM Data Management* magazine, change your address, or make other updates to your information, please go to ibm.com/dmmagazine. For instant access to the *IBM Data Management* magazine digital edition, visit ibmdmmagazinedigital.com/dmmagazine. For information about reprints, please send an e-mail to customerservice@tdagroup.com.

IBM and the IBM logo are registered trademarks of the International Business Machines Corporation and are used by TDA Group under license.

Material published in *IBM Data Management* magazine copyright © 2010, International Business Machines. Reproduction of material appearing in *IBM Data Management* magazine is forbidden without prior written permission from the editor.



PRESIDENT

Paul Gustafson

VICE PRESIDENT, STRATEGY AND PROGRAMS

Nicole Sommerfeld

VICE PRESIDENT, EDITORIAL DIRECTOR

Debra McDonald

VICE PRESIDENT, CONTENT SERVICES

Paul Carlstrom

Printed in the U.S.A.



We do our best to assume a calm, unruffled outward appearance, but putting together a magazine takes a large, talented, and opinionated team. And the editorial offices of *IBM Data Management* magazine are occasionally the scene for pointed discussions about both topic and approach.

In fact, the theme of this issue—databases that push the limits of size and speed—was the subject of much debate. Several folks argued that the topic was too esoteric and that the issue would amount to a sideshow of the exotic and the hyperspecialized: “Why spend time on examples that only a few of our readers will ever encounter?”

Ultimately, we decided that this tour of the outer limits was warranted, if only due to the persistent tendency of information technology to make commonplace what once resided only on the fringes of possibility. But the argument haunted us as we assembled the issue you hold in your hand. At every turn, we pushed harder, asking ourselves how the exciting vistas that we were seeing related to the everyday challenges that data professionals face.

For me, our exploration of the very large and very fast illuminated some of the most basic principles of data architecture and project management. In his Data Manager column, for example, Merv Adrian interviews large-database maven Richard Winter, eliciting an analogy that drives home the simple virtues of planning ahead. And our cover feature provides ideas on how to think about distributed data that ring true no matter what size the project.

A few pages later, a story on the technology behind the in-memory database IBM solidDB illustrates the importance of precisely defining one's terms. In discussions about extreme speed, a high transaction rate does not necessarily indicate low latency—a critical distinction. And while we're going behind the scenes, Lester Knutsen gives us a look at some of the techniques that won the most recent Fastest Informix DBA contest, and an object lesson in the value of sweating the small stuff.

In retrospect, we aimed for the stars with this issue (and we got there, if only with our back-page story about a radio telescope project that will generate a truly mind-boggling amount of data). We hope that we were able to turn and reflect some of that light back to the ground that we all tread on. Give me a shout at editor@tdagroup.com and let me know what lessons you learned, and what you'd like to see us tackle in the future.

Thanks for reading,

Cameron Crotty
Editor

FREE Resources for DB2 Users!

Join IDUG:
The Worldwide
DB2 User Community

- FREE videos with the best DB2 technical presentations
- FREE access to prior year's conference presentations
- FREE access to technical articles
- Get answers to your questions on the DB2-L listserv
- Information about IDUG's global conferences and regional events
- Calendar of Regional User Group meetings
- 45% discount for books from IBM Press
- and much more...

Sign-up is FREE at
[www.IDUG.org!](http://www.IDUG.org)

IDUG's worldwide user community represents more than 11,000 members in more than 100 countries around the globe. Dedicated to users of IBM's DB2 family of products and the tools that support them, IDUG helps DB2 users improve their professional efficiency and their organization's return on investment from DB2.



IDUG 2010 - North America

The Ultimate DB2 Conference

May 10-14, 2010
Tampa, Florida

Your **only** source for independent, unbiased, and trusted DB2 information.
If you are going to attend only one conference, this is it!

- The **most** DB2 technical sessions of any conference
- The **best** DB2 technical sessions in the world
- DB2 certification -> no additional charge
- Fun networking activities
- Meet fellow DB2 users and leading DB2 consultants
- Access IBM experts and developers
- **NEW** - IBM hands-on labs -> no additional charge
- **NEW** - IDUG Mentor program: Leverage your skilled assets to secure long-term success

Learn more at [www.IDUG.org!](http://www.IDUG.org)



IDUG
The Worldwide DB2 User Community

IDUG Headquarters • 401 N. Michigan Avenue, Suite 2200 • Chicago, IL 60611-4267
Tel: +1.312.321.6881 • Fax: +1.312.673.6688 • E-mail: idug@idug.org • Web: www.IDUG.org

The International DB2 Users Group (IDUG®) is an independent, not-for-profit, user-run organization whose mission is to support and strengthen the information services community by providing the highest quality education and services designed to promote the effective utilization of the DB2 family of products.

The DB2® Data Servers include DB2 for z/OS; DB2 for Linux, UNIX, Windows; DB2 for i; DB2 Server for VSE and VM; DB2 Express; and DB2 Everyplace.

Smarter technology for a Smarter Planet:

How to manage thousands of things you can't touch.

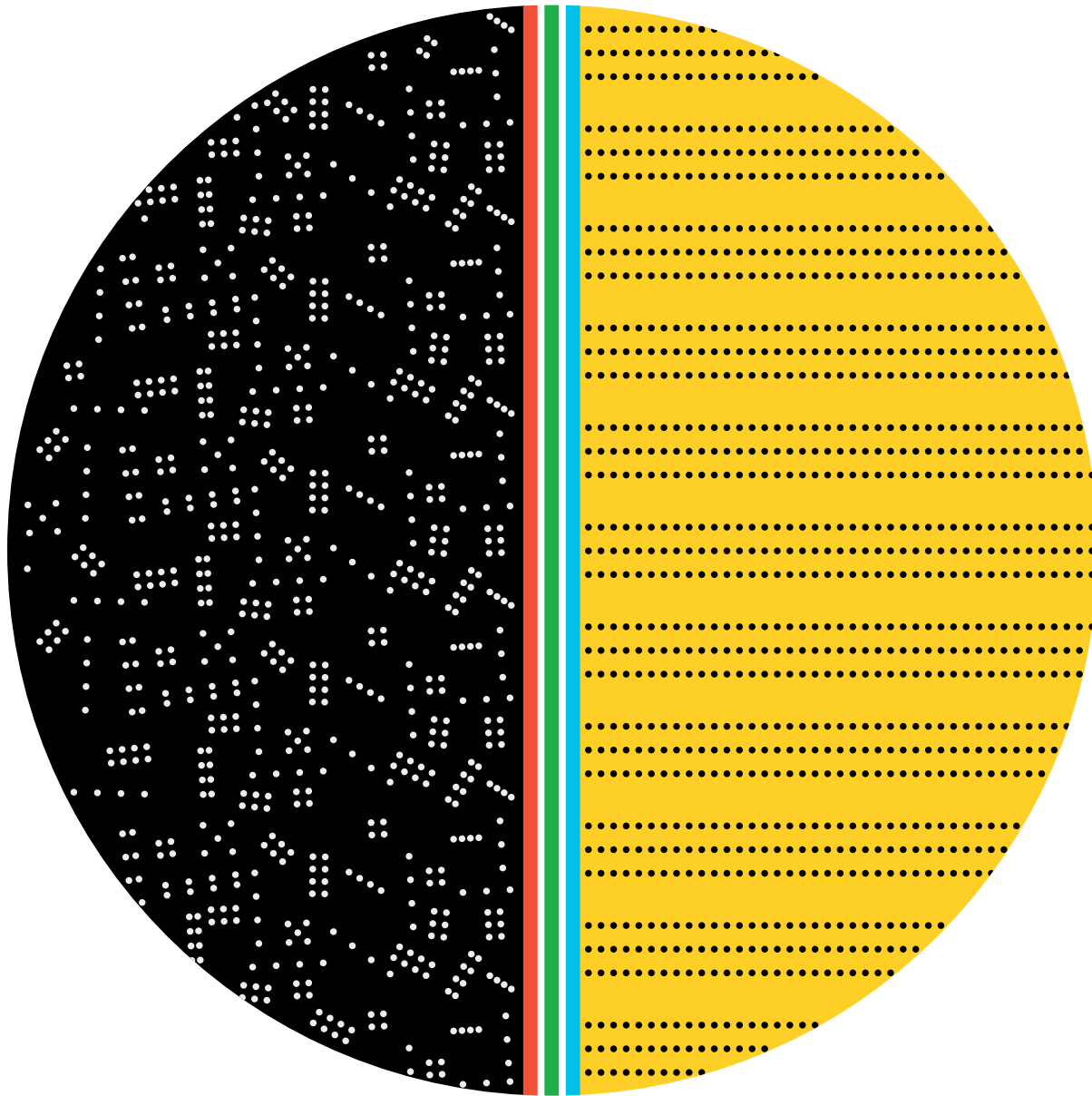
As virtualization gains momentum, many companies are finding out the hard way that virtual image sprawl can be just as complex and overwhelming as the physical server sprawl it was meant to solve.

Forty years ago, IBM pioneered virtualization. Today, IBM can help you manage, simplify and even automate your virtual environment with a broad range of solutions designed to give you visibility and control over all of your virtual resources—servers, storage, applications, etc. So you can provision and reconfigure resources in minutes instead of days, driving up efficiencies and setting the stage for new delivery models like cloud computing.

Our open approach to virtualization has helped customers reduce capital and operating costs by up to 30% and is becoming an essential building block of a smarter, more dynamic infrastructure.

A smarter business needs smarter software, systems and services. Let's build a smarter planet. ibm.com/virtualize





IBM Acquires Market Leader in Database Activity Monitoring

Guardium technology helps identify unusual access patterns

On November 30, IBM announced that it had acquired Guardium, a leader in real-time enterprise database monitoring and protection. Guardium automates regulatory compliance tasks to help safeguard data, monitor database activity, and reduce operational costs.

Designed for cross-platform environments, Guardium technology identifies patterns and anomalies in data access and usage, which allows organizations to maintain the integrity of their data and turn it into a strategic business asset. The monitoring capabilities of the technology also detect fraud and unauthorized access via enterprise applications such as an organization's enterprise resource planning (ERP), customer relationship management (CRM), or data warehousing solutions.

The tools and technology that IBM has added to the IBM Information Management portfolio will enable IT organizations to maintain trusted information infrastructures by continuously monitoring access and activity to protect high-value databases

against threats from legitimate users and potential hackers. Centralized and automated controls for all major platforms will also help streamline compliance processes for ever-changing industry and government mandates. "Organizations are grappling with government mandates, industry standards, and business demands to ensure that their critical data is protected against internal and external threats," says Arvind Krishna, general manager, IBM Information Management. "This acquisition is another significant step in our abilities to help clients govern and monitor their data, and ultimately make their information more secure throughout its life cycle."

The combination of IBM and Guardium technology will help organizations safely realize the promise of business analytics and use trusted information to drive smarter business outcomes.

MORE INFORMATION

www.guardium.com

IBM Information On Demand EMEA Conference 2010 in Rome

Mark your calendar for the IBM Information On Demand EMEA Conference 2010, which will be held May 19–21 at the Marriott Park Hotel in Rome, Italy. This event is your best opportunity to learn how to better optimize your data management resources, reduce costs, manage risk, improve customer insight, and increase operational efficiency. A comprehensive technical and business leadership program will provide you with an unrivalled forum to share best practices, network with your peers, and learn the latest about data management innovations such as cloud computing, which are helping to drive an information-led transformation.

Why you should attend

- ▶ More than 2,000 attendees from across Europe, the Middle East, and Africa
- ▶ Thought-provoking general sessions from leading industry commentators and IBM executives
- ▶ Technical and business leadership tracks
- ▶ Compelling speakers, including business luminaries, analysts, and customers
- ▶ One-to-one meetings with more than 50 IBM senior executives

- ▶ In-depth solution, industry, technical, and subject matter insight from experts
- ▶ The Expo Solution Center, showcasing offerings from IBM Business Partners and the pan-IBM organization
- ▶ Exclusive Business Partner Program on May 18
- ▶ World-class technical training, free certification testing, and hands-on labs
- ▶ Extensive networking opportunities

▶ MORE INFORMATION

ibm.com/software/europe/data/conf

DB2 First DBMS to Achieve VMware Ready Certification

Data server and virtualization technology work together to maximize resource utilization

VMware, Inc., a provider of virtualization solutions, recently announced it will expand the scope of its VMware Ready program to allow middleware and application software to qualify for the VMware Ready logo. IBM DB2 was the first and only database management system to be certified as VMware Ready at the time of the program's expansion.

The VMware Ready certification simplifies the purchase and deployment process for customers and prospects by signifying that qualified applications deliver outstanding performance and reliability when deployed on VMware vSphere. Nearly 1,000 of the most widely used applications from independent software vendors (ISVs) support the VMware vSphere platform.

"We are living in an age in which businesses must have access to vast quantities of information as a tool for making smarter and faster decisions, at a lower cost," says Bernard Spang, director of Information Management Strategy at IBM. "Through the VMware Ready program and the long-standing collaboration between our two companies, our mutual clients can confidently enjoy the cost and energy savings that both IBM DB2 software and VMware virtualization technologies deliver."

Parag Patel, vice president of alliances at VMware, adds: "Our partner ecosystem is among the largest in the industry, and having a wide selection of proven applications for VMware vSphere sets us

apart from everyone else. The VMware Ready logo gives participating ISVs a powerful sales advantage and the ability to inform customers that their applications leverage the TCO savings and flexibility of the VMware platform. And, customers will have confidence and peace of mind knowing their applications are ready for the VMware platform. From the top of the technology stack to the lowest layers, customers have more and more choices with VMware."

▶ MORE INFORMATION

www.vmware.com/partners/vmware-ready



IDUG North America Conference to Offer Hands-on Labs

Explore features and try new software

The International DB2 Users Group (IDUG) is expanding its 2010 North America conference lineup with hands-on labs that will give attendees deeper insight into the newest IBM DB2 features and tools. The labs will be included with regular IDUG conference registration and will be offered throughout the four-day conference, held May 10–14 in Tampa, Florida.

“The instructors will take the students through a variety of tasks that get right to the point of demonstrating the featured software,” says Fred Sobotka, a member of the IDUG Conference Planning Committee. “If you’ve ever wanted to explore new DB2 features or try out a product, but didn’t have the time to go through the hassle of obtaining the software and setting it up, these labs cut to the chase and quickly deliver the knowledge and experience you expect from an IDUG conference.”

In addition to the labs, IDUG North America provides a wealth of technical education and networking opportunities, including more than 100 hours of peer-reviewed technical sessions for DB2 for z/OS and DB2 for Linux, UNIX, and Windows; expert panels of top DB2 consultants and IBM engineers; free certification testing for dozens of IBM Information Management exams; and an expo area filled with the world’s top third-party DB2 tool vendors and experts.

➤ MORE INFORMATION

www.idug.org/NA

IDUG Launches Mentor Program

Special discounts available to help train new DB2 users

The International DB2 Users Group (IDUG) has announced an IDUG Mentor program to help long-term members train new IBM DB2 professionals. Anyone who has attended five IDUG conferences in the past 10 years is invited to apply to become an IDUG mentor.

IDUG mentors will receive special recognition at this year’s IDUG North America conference, May 10–14 in Tampa, Florida. At the conference, mentors will meet with DB2 developers and presenters and participate in discussions on how they are working to prepare the next generation of IBM DB2 professionals. Mentors also will be eligible to apply for a special price to bring a co-worker to the conference.

“We have been hearing the same thing over and over from our members—

management wants us to have a succession plan, but they are not providing additional training because of the economy,” says Michael McBride, IDUG president. “Our board has decided that this is something IDUG needs to help with as part of our support of the worldwide DB2 user community. So we are offering a special discount to allow IDUG mentors to sponsor and bring the next generation of DB2 professionals as a first-time conference attendee.”

Full details of the program and an application to become an IDUG mentor are on the IDUG Web site at www.idug.org.

➤ MORE INFORMATION

www.idug.org

DISCOVER NEW AND POPULAR DB2 9 TRAINING OPTIONS

Get on the path to IBM DB2 9 skills improvement in 2010 by enrolling in an IBM training course. A number of flexible training options are available, including traditional classroom, instructor-led online, and self-paced virtual classroom courses, so you can choose the course and the delivery method that work best for you. To simplify your selection, view the online training paths in your interest area.

➤ MORE INFORMATION

ibm.com/software/data/education/roadmaps.html

MIGRATION FROM ORACLE TO DB2: A NEW SERVICES SOLUTION OFFERING

Do you have IBM DB2 as well as a sizable Oracle footprint, and are you looking to jump-start your migration to DB2? If so, you might benefit from a new services solution offering from IBM. This total solution approach to migration is based on proven methods and best practices, and it leverages existing IBM asset tools to accelerate migration. This offering consists of four phases:

- Migration assessment
- Migration strategy
- Database and software conversion
- Testing

➤ MORE INFORMATION

http://download.boulder.ibm.com/ibmdl/pub/software/data/sw-library/services/DB2_Competitive_Migration_Service.pdf



System z Academic Initiative Reaches More Than 600 Schools

Number of participating institutions continues to grow

IBM recently announced that the number of schools taking part in the System z Academic Initiative has surpassed 600 worldwide. Since the program's inception in 2004 with 24 colleges in the U.S., the multimillion-dollar program to help colleges, universities, and high schools build students' mainframe computer skills has expanded to more than 60 countries.

More than 50,000 students worldwide have taken part in IBM enterprise systems education. For educators and students, the initiative provides a comprehensive enterprise systems curriculum, faculty workshops, resume posting services, access to industry experts, and special tests and contests. Mainframe skills are in demand from banks, government agencies, airlines, retailers, and others in an interconnected world, and IBM System z is the gold standard for superior reliability and management of high volumes of computer transactions.

Schools such as Marist College, a private college in Poughkeepsie, New York, have created programs as part of the System z Academic Initiative. The new z/OS IMS Application Programming Certificate at Marist provides an introduction to both System z and IMS. The program, which is a series of three classes, introduces students to z/OS terminology and concepts, provides an overview of the IMS Database Manager and IMS Transaction Manager environments, and provides students with techniques for writing application programs in those environments.

"For more than 45 years, IBM has continued to innovate and protect our customers' investments in the mainframe," says Tom Rosamilia, general manager of System z at IBM. "By investing with more than 600 schools to train the next generation of IT leaders, we ensure our customers not only have a supply of critical skills, but continue to realize unmatched returns on the unique strengths of their System z platform."

MORE INFORMATION

ibm.com/university/systemz

Enhance Your Career with DB2 Certification

Earning an IBM industry-recognized professional certification is a proven way to demonstrate your technical skills and abilities. Plan your 2010 IBM DB2 9 certifications with the help of Certification Roadmaps. View available exams and prerequisites to plan for success. New certifications are available for DB2 9.7, including IBM Certified Database Administrator DB2 9.7 for Linux, UNIX, and Windows (Exam 541) and IBM Certified Application Developer DB2 9.7 for Linux, UNIX, and Windows (Exam 543). Advanced DB2 9.7 certification will also be available in 2010.

MORE INFORMATION

ibm.com/software/data/education/cert-roadmaps.html

Online Exclusive: Gauge SQE Use in DB2 for i 6.1

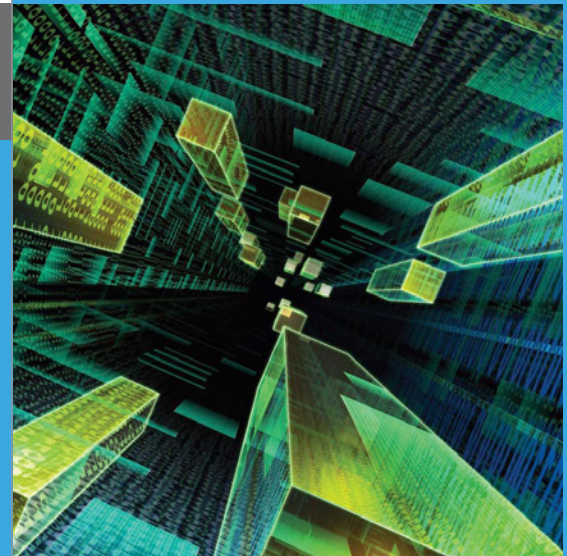
Explore the online *IBM Data Management* magazine for more great content, including a recently published article by IBM's Scott Forstie about SQE use in DB2 for i 6.1. This article explains how to collect an SQL Performance Monitor, understand which SQL queries are using SQE, and understand why other SQL statements continue to use CQE. SQE is the strategic query engine on DB2 for i, and it provides better performance, scaling, and tools such as the SQL Plan Cache.

READ THE ARTICLE

tiny.cc/DB2fori

Explore more great online content:

ibm.com/dmmagazine





Stuart Litel
is president of the
International
Informix Users

Group (IIUG; www.iiug.org/president), CTO of Kazer Technologies (www.kazer.com), an IBM Gold Consultant, member of the IBM Data Champion Inaugural 2008 class, and recipient of the 2008 IBM Data Professional of the Year award.

To IOD or IIUG?

Broad-spectrum view or
all Informix, all the time

Users often ask me what seems like a simple question: “If I can attend only one conference, should I go to IBM Information On Demand (IOD) or the IIUG Informix Conference?” When I hear this question, I have a two-word reply ready to go: “It depends.”

First, let me give you my review of this past October’s IBM IOD conference in Las Vegas. As usual, IBM put on a very nice show over four great days at the Mandalay Bay Resort and Casino. The conference featured two IBM software brands: what is now called Business Analytics and Performance Optimization as well as Information Management. Former General Manager of Information Management Ambuj Goyal told the crowd that he would be taking over the new software brand of Business Analytics and Process Optimization; he also introduced the new General Manager of Information Management, Arvind Krishna. We were all treated to the best of Informix, DB2, Cognos, Content Management, InfoSphere, solidDB, and much, much more.

This event is not to be missed if you want the full effect of the IBM Information Management as well as Business Analytics and Performance Optimization brands. Here, Informix isn’t the main subject, but you get to hear about it in the context

of all the parts of the IBM Information Management as well as Business Analytics and Process Optimization portfolios.

How does this compare to the IIUG Informix Conference? The IIUG Informix Conference is dedicated to and centered on one thing: Informix. Of course, we feature IBM software products such as Cognos

The IIUG Informix
Conference is
dedicated to
and centered
on one thing:
Informix.

and solidDB, but the focus is on how these solutions can be used with Informix.

This is the third year of the IIUG conference, and it includes more than 75 dedicated Informix sessions to help you understand how to make your Informix database run faster, better, and more reliably. You will hear from the “best of the best,” including both the architects at IBM who actually create Informix and users just like you from around the world (the 2009

event was attended by users from more than 25 different countries in North and South America, Europe, and Asia).

This year’s event will feature a keynote from General Manager Arvind Krishna, and will be attended by IBM vice presidents, directors, and other key Informix individuals from IBM. This conference is truly a learning experience, and the primary focus is all about learning and education. (OK, the parties are not bad, either—you will meet users from around the world and make lifelong Informix friends.)

So which event would make you go to your boss, bang at the door, and say, “I must go!?” Well, if you want an event centered around Informix, the IIUG conference is your choice. If you are looking for a wider view of IBM Software Group products that includes Informix, IOD is the place for you.

The IIUG conference will take place in Overland Park (Kansas City), Kansas, on April 25–28, 2010 (see www.iiug.org/conf for more details). If you want to join us in April at the completely renovated Overland Park Marriott, register before March 15, 2010, and you’ll pay less than \$500 (price available to all IIUG members, plus an additional \$100 off if you use the registration code: MAGDISC100). See you in Overland Park! *

The Next DB2 for z/OS

First glimpse into the future



David Beulke (dave@davebeulke.com) is president of Pragmatic Solutions, Inc. (PSI), a

training and consulting company that specializes in designing and improving SQL, application, and system performance on DB2 for Linux, UNIX, and Windows, and z/OS. He has experience in the architecture, design, and performance tuning of large data warehouses and OLTP solutions. He is also a former president of IDUG.

During the last several months, at the International DB2 Users Group (IDUG) and IBM Information On Demand (IOD) conference sessions, much information has surfaced regarding the next version of IBM DB2 for z/OS—which for now is being called “DB2 X.” Even though this new version is still in development and no release date has been set (late 2010, maybe?), several new powerful features are very interesting.

First, the traditions of performance, scalability, reliability, availability, serviceability, and superior security on DB2 for z/OS will continue. DB2 X looks to enhance these capabilities by including additional resources for service-oriented architecture (SOA), .NET, and Java workloads, as well as improved support for software from SAP, Oracle PeopleSoft, and others.

DB2 X will continue to make it easier to develop state-of-the-art applications. Today, DB2 for z/OS gives developers a host of options: they can use simple or complex SQL, take advantage of XML with pureXML or xQuery, and accelerate Java applications with the System z Integrated Information Processor (zIIP) engines. DB2 X also continues to enhance its support of SOA frameworks such as .NET, Java, and cloud architectures, and will interface with Hadoop processes to handle new application ideas and requirements.

After listening closely to customers, IBM is making several enhancements that will enable developers to create an even wider variety of applications for the z/OS environment. In DB2 X, you’ll see SQL improvements that enhance data warehousing applications, including moving sums, moving averages, better query optimization, and more parallelism.

Beyond the SQL improvements, DB2 X also offers pureXML tools for leveraging vast quantities of unstructured data, as well as the ability to create temporal queries and to timestamp with time-zone-improved sensitivity. These tools and capabilities enable developers to create more capable applications and give data increased context.

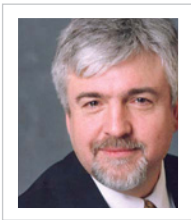
These enhancements also make it easier to port applications from databases such as Oracle or SQL Server to the DB2 for z/OS environment. We are seeing the number and pace of migrations to DB2 for z/OS pick up because of performance, reliability, and overall security features. The combination of the mainframe and DB2 for z/OS is increasingly becoming the internal “cloud computing” solution of choice for performance- and security-conscious corporations.

In addition, the newly announced IBM Smart Analytics Optimizer is a new blade server attachment to DB2 for z/OS. It demonstrates up to 5 to 10 times performance improvement for existing data warehouse

queries according to IBM testing benchmarks. The Smart Analytics Optimizer is a network-attached blade appliance that enables dynamic data warehousing and operational business intelligence through linear query scaling without any changes to the existing applications. This special-purpose appliance caches, cubes, and optimizes your data so that your existing data warehouse and business intelligence workloads receive dramatic performance and response time improvements. The Smart Analytics Optimizer is currently in beta in case you or your company might be interested in getting involved and finding out how this solution will advance your DB2 environment.

Within the IDUG and IOD conference sessions, there were great discussions of even more features for DB2 for z/OS. Additional information will become available as IBM development continues. Watch my blog (www.davebeulke.com) and several blogs featured within the IDUG Web site (www.idug.org/blogs.html, www.planetDB2.com) and IBM developerworks (ibm.com/developerworks/mydeveloperworks/blogs/) for further details as they become available. Be sure to watch and budget for the next IDUG conference in Tampa, Florida, in the spring, where more sessions will further detail all the great features that are coming in DB2 X for z/OS and the DB2 family of products. *

Exploring the Extremes of Database Growth



Merv Adrian is principal of IT Market Strategy, a research consultancy that analyzes software trends and advises leading IT firms on market strategy, the competitive landscape, and go-to-market execution issues.

An interview with Richard Winter

Richard Winter has been watching the limits of database size for more than a decade. WinterCorp, his independent consulting firm, helps enterprises optimize the value of multi-terabyte databases and data warehouses throughout their life cycle. Richard's expertise is widely recognized, so we were happy to have the opportunity to chat with him recently to discuss growth patterns, skill sets, and technology developments.

Richard, you've been tracking the largest database installations in the world for some time, and in the past decade, the boundaries have been rewritten continually. Just how rapid has the growth been?

The rate of growth of very large databases, especially data warehouses, has been truly remarkable over the last 10 years. In WinterCorp TopTen surveys from 1998 to 2005 and in anecdotal studies since then, we have observed that the size of the largest data warehouse triples about every 2 years. In fact, we have gone from terabytes of data

to petabytes of data—a factor of 1,000—in the present decade.

We now have many more tera-scale data warehouses than we used to. They were a rarity early in the decade and now there are hundreds or thousands of them. But, it also means that a 5 terabyte data warehouse is no longer “very large.” When WinterCorp talks about VLDB [very large databases] or VLDW [very large data warehouses], we are still talking about the very largest systems: the 10 or 20 in each arena that tower above all the others. Today, in terms of volume, those systems are in the hundreds of terabytes or petabytes of data.

Have the practitioners of VLDB proliferated as well, or are there only a few truly experienced teams out there? Do the required skills change at the top end of the volume chart?

The skills to plan, manage, design, and implement those systems—the systems that define the frontiers of scalability—are still in

high demand and short supply. The issues of database scale are not only about how much data you have. It is also about workload. And, it is about complexity. If the schema or the queries are very simple, data volume is not as big a problem. But if the schema is complex, the queries are complex and unpredictable, there are many concurrent users, and there is continuous update with low latency, then you have a really challenging problem in database scalability. Those problems have not changed. Only the location of the frontier has changed.

What management techniques separate stewards of the largest databases from the rest?

There are many differences in managing on a large scale. One of the most fundamental differences is that you have to plan ahead—and measure ahead—more. You have to proactively manage performance.

An analogy I like is walking uphill. If you are going for a walk in the Berkshire Hills of Massachusetts (elevation about 2,500 feet) on a summer afternoon, you don't need much planning; and, you don't need any specialized skills or equipment. If you run into some bad weather, you can just turn around and walk back to your car or your home. But if you are climbing one of the Himalayan peaks (elevation greater than 20,000 feet), you may still be just walking uphill (most of the time) but you need to approach the problem very

differently. You must anticipate how it's going to be different at higher altitudes and prepare in advance. You can't wait until you are up above the tree line to figure out how cold or windy it is going to get.

Similarly, with these very large databases, you have to assume you may encounter problems greatly amplified by scale. And you can't wait until the database is loaded and the applications are built to find out what most of those problems are. You need to plan ahead, measure performance, measure scalability, and quantitatively estimate the effects of design decisions or requirement changes. In general, you need to do all you can to anticipate, prepare for, quantify, and measure problems and solutions in advance.

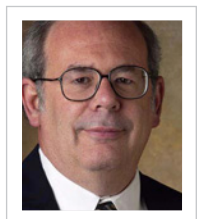
To what degree do you see large volumes of data being used outside of traditional DBMSs in file stores? Are there best practices for combining these separate stores with DBMSs using new tools such as Hadoop and MapReduce?

Newer techniques such as MapReduce are an important and exciting direction for data analysis. They provide a framework for using procedural analysis and widely used programming languages and techniques on highly parallel systems. As a result, they are being used on data volumes 10 to 100 times greater than the volumes we see in our largest data warehouses.

At the same time, using MapReduce in combination with data warehouse technology is opening the door to interesting new approaches. Several of the analytic database start-up vendors, including Aster Data, Greenplum, and Vertica, have provided capabilities along these lines. Several other vendors have announced or are developing such capabilities.

At a recent analyst event, Rod Smith, vice president of Emerging Internet Technologies at IBM, made it clear that IBM is not ignoring the MapReduce/Hadoop phenomenon. He talked about field projects with customers that investigate the use of Jackal for distributed shared memory, user-defined functions, the Pig language for data analysis with Hadoop, and the exporting of results into feeds and XML. He also showed a visual front end designed to bring the advantages of Hadoop to less-technical business users. What guidelines do you think users should apply here?

Hamid Pirahesh, an IBM Fellow, has also been doing some fascinating work in this area. It is important to appreciate that SQL and MapReduce have different strengths. One of the strengths of SQL is non-procedural selection, retrieval, aggregation, and manipulation of data. The best modern, highly parallel SQL engines will optimize basic operations on data with great sophistication, making data-dependent decisions based on the size, structure, cardinality, and skew of structured data in tables. It would be a great deal of effort—and a wasted effort today—for users to replicate these capabilities in procedural application code in MapReduce, although computer science efforts to do so may prove useful as they mature. So, when there is an integrated facility, people should aim to have the SQL engine do as much work as possible before dropping into the procedural domain of MapReduce. Avoid reinventing the wheel—use MapReduce primarily to do the things that SQL doesn't do well or simply. ✱



“The skills to plan, manage, design, and implement those systems—the systems that define the frontiers of scalability—are still in high demand and short supply. The

issues of database scale are not only about how much data you have. It is also about workload. And, it is about complexity.”

—Richard Winter, *WinterCorp*

1011011000 01

extreme

databases:

Learning from
life on the
database edge

11000 11010 00110101

1011011000 011

e n n e

the biggest and fastest

By Eric Ahrendt

Calling something big or fast immediately begs the question, “Compared to what?” A “big” database for a small company is dwarfed by a national data repository growing by 28 petabytes per year; a “fast” database that processes transactions for an e-commerce site is slow compared to one that delivers access times measured in milliseconds in order to automatically execute stock trades.

But even if your company isn't in the running for the biggest or fastest database on the planet, the lessons in administering such databases may be applicable to your environment. It's a sure bet that the trends in this realm are going to filter down to databases of all sizes.

“...today, a database must be in the multi-petabyte range to be considered extreme.”

—Dr. Robert Hollebeek
Professor of Physics
University of Pennsylvania

Defining extremely large

As the total amount of data created continues to grow, public and enterprise databases are expanding to hold it. Just four years ago, WinterCorp¹ identified the world's largest databases, which were data warehouses measuring over 100 TB. A Yahoo! database was the first system in the 10-year history of the program to surpass the 100 TB mark.

But now, in a world with a lot more digital information to store, what defines an “extremely large” database? There's no standard definition, and size alone is no longer the only criterion—manageability is also a factor. One workable definition comes from Dr. Robert Hollebeek, professor of physics at the University of Pennsylvania, who co-founded the National Scalable Cluster Project and has won several national awards for work in distributed clustered systems and data mining. Hollebeek says that five years ago, a multi-terabyte database would have qualified, but today, a database must be in the multi-petabyte range to be considered extreme. “Another definition would be a database whose index can no longer fit in physical memory—even the terabyte memory of a supercomputer or cluster of machines,” he says. A database requiring an index that large is extreme and creates significant problems related to performance and database administration.

¹ 2005 Winter TopTen Program Award Winners, WinterCorp, September 14, 2005, http://www.wintercorp.com/VLDB/2005_TopTen_Survey/2005TopTenWinners.pdf.

Hollebeek says a database also qualifies as extremely large when the total mass of hardware required to house the data becomes problematic: “When you have thousands of disk units or a roomful of racks on parallel machines, it becomes difficult to manage.”

IBM Information Champion Manuel Gomez Burriel of the Confederacion Española de Cajas de Ahorros (CECA), a Spanish confederation of savings banks, agrees that manageability can separate a garden-variety “large” database from an extreme one. “Common administration tasks can become impossible to complete in the available time window,” says Gomez. Recovering from a corrupted database may take hours when only minutes are available. Performance can suffer because the database is too large for any significant part of it to fit in an in-memory cache. And simply responding to application requests for data can come with unacceptable costs in CPU cycles.

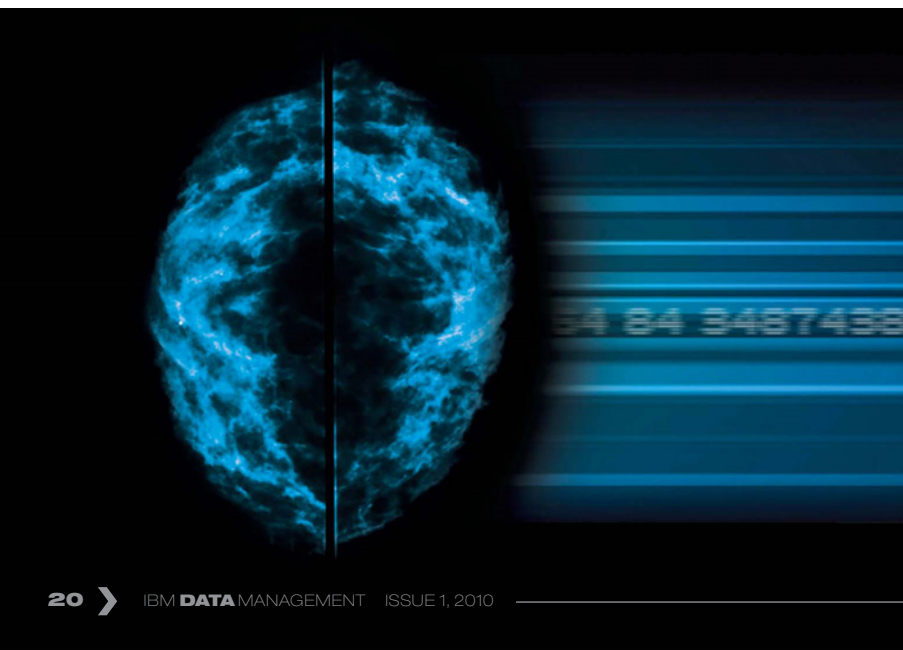
Portrait of a database

Looking closely at a single very large database yields lessons in data management that can apply to other large and not-so-large databases. Hollebeek served as technical lead on the National Digital Mammography Archive (NDMA), a system designed to include a database growing by 28 petabytes per year. Funded by the U.S. National Institutes of Health, NDMA established a distributed grid of systems for medical records and image storage. The system stored mammography scans, MRI scans, and related files that made up a “case,” each of which could be as large as a gigabyte—and there were millions of them.

Aside from solving problems in storing and accessing massive amounts of data, NDMA had to deal with issues related to siloed data stored on systems that were geographically distributed—a common problem among global enterprises. To link with the four research hospitals participating in the project, NDMA installed encrypted, secure lines and a “point of presence” in each hospital, which was hardware that encrypted files and used a special network protocol for efficiently sending large blocks of data.

“Our project was very large-scale, and we couldn't lose any of the medical data. We needed rock-solid, highly reliable technology that was also really fast and parallel-capable because the structure was based on building clusters of parallel machines,” Hollebeek says. “It had to be fault-tolerant because it would be unacceptable for index tables to crash or disappear.”

NDMA used IBM DB2 Parallel Edition software to store the database index. NDMA stored the actual image data in flat file databases on parallel disk arrays, where it was managed by the native file system for the operating system, which in this case was Linux.



Lessons from NDMA

From his work on NDMA, Hollebeek passes on a few recommendations for dealing with extremely large databases connected by a wide area network (WAN):

- ▶ Pay attention to the problem of sending large amounts of data over the network, whether that's the Internet or a private network. Look for ways to do that efficiently, such as installing points of presence at sending and receiving locations or using a protocol that sends large (many megabyte) blocks.
- ▶ Don't change the format of the data you receive. Lossless compression is fine, but reformatting and storing received data for a small percentage gain in space in a large database will cause more issues than it's worth. In the case of medical data, for example, it's especially serious if you make a mistake when changing the format and the data becomes unusable.
- ▶ Database performance plummets once index tables don't fit in memory anymore, so scale up memory as far as possible to hold tables. Then take advantage of any parallel structure in the data to effectively add more capacity using clusters. If that fails, make an index for the index.

Defining extremely fast

Like the definition for extremely large, the definition for extremely fast changes as technology continues to push the limits of what's possible. "We usually measure database speed in something like transactions per second, and maybe a million transactions per second is fast. But with new technologies, the term 'transaction' defines only a single, fairly conventional class of database access," says Carl Olofson, IDC research vice president, Application Development and Deployment. "In the future, there will be other classes that blur the distinction between database and memory. At that point, pulling data may involve some kind of redirected memory access as opposed to actually binding to a database and initiating a transaction. And that can obviously be very fast."

Defining fast in a broader sense, Olofson says that "the aim of any system is to ensure that the database doesn't slow down the application, so you can return data as quickly as the application runs."

Gomez takes a similar tack by defining extremely fast as "fast enough to deliver information according to the SLAs [service level agreements] agreed upon with clients." He adds that "the fastest databases have direct access to a piece of information, in memory, if possible. One of our payment-system applications uses IMS with a Fast Path Solution

When microseconds matter

Bolsa de Comercio, Chile's stock exchange in Santiago, Chile, has very strict service level agreements (SLAs) for both responsiveness and recovery. The IT team added IBM solidDB Universal Cache as a front end to their existing data repository, built on Microsoft SQL Server, to address the high numbers of transactions per second coming from IBM WebSphere Low Latency Messaging software (MQ LLM) during times of peak trading. Since then, all transactions consistently complete in microseconds, and the exchange has the ability to transparently recover from system failures in less than a second.



and an enterprise storage subsystem to deliver a less than 20-millisecond response time per transaction, and it's accessing up to 14 databases for client information."

"...the ability of a system to deliver that kind of speed can mean the difference between success and failure in managing these accounts."

—Carl Olofson

IDC Research Vice President
Application Development
and Deployment

A second delay is too long

Whereas the financial systems behind automated teller machines (ATMs) used to be considered fast, they're now also-rans. "These days, you may have to wait a second or so to see your account balance," says Olofson. "That used to be great; now it's average." Instead, when experts talk about high-speed database interactions today, they're talking about telecom systems where in the course of connecting a call, a system instantaneously looks up the customer's account to determine what class of service they have so the system knows how to route the call and what features should be made available—all in an environment like the wireless world where anyone's account can change at any time.

Another example of an application that requires extreme speed is algorithmic trading driven by portfolios in the financial services industry. "A firm may have hundreds of accounts, each with a portfolio that's slightly different, and each therefore has different rules that govern what kinds of trades can be made," says Olofson. "These rules must be applied within milliseconds of a price coming over the wire, and the ability of a system to deliver that kind of speed can mean the difference between success and failure in managing these accounts."

Such extreme speed requires back-end databases that are highly responsive, which typically involves a tiered system of databases. In many cases, the back-end database is actually a mainframe database such as IMS, IBM's mainframe-based hierarchical DBMS. An in-memory database caching capability acts as a front end to a back-end database in these tiered solutions. At the forefront of investing in such tiered solutions are the customers that need extremely fast systems.

Speed-boosting technologies on the horizon

The demand for speed never lets up, and vendors continue to develop new ways to accelerate database performance. One tactic that's gaining popularity is to attack what is usually the slowest part of the data transfer chain: the hard drive. In-memory caching solutions such as IBM solidDB move the database off of relatively slow hard drives into relatively fast RAM, dramatically improving response time. For a more in-depth look at solidDB, see "solidDB and the Secrets of Speed," this issue.

Another promising solution is solid-state drives (SSDs) or flash memory, which now is a sandwiched tier between main memory and disk memory, but could become the principal storage medium for databases as its cost comes down. (See "Unleashing the Value of Solid-State Drives for DB2 Workloads" in this issue for a discussion on how to maximize the performance of DB2 with SSDs.)

Along with the shift in hardware away from disk-centric and toward memory- and processor-centric data management, flexible technologies are being developed to address different database workloads. Instead of the conventional way of storing data as records that then become rows in a table, data managers are using columnar databases and data pools pointed to by matrices of indexes that offer a high degree of flexibility for internal storage. From the perspective of the application accessing the data, such a repository looks like a conventional relational database, but it can be expanded to work as an object database, an XML database, a multi-value database, or a multi-dimensional database.

Finally, some developers are taking a completely different approach called "stream processing," which focuses on processing and analyzing data as it is collected, instead of waiting until it has come to rest in a database. IBM InfoSphere Streams takes this approach, making it possible to monitor simultaneous data streams and provide nearly instantaneous analysis that is continually refined as new data becomes available. For a closer look at InfoSphere Streams working in the real world, check out "Smarter is: Boosting the IQ of Galway Bay," in *IBM Data Management* magazine Issue 3, 2009.

It's clear that in the realm of extreme database size and speed, customer needs continue to drive technology advances. No matter how much data you need to store and retrieve, or how fast, companies are working on products to make it possible. And just as in language, fashion, and art, what's extreme today will be run-of-the-mill tomorrow. *

Eric Ahrendt writes on technology for a range of Fortune 500 companies.

RESOURCES

IBM DB2: ibm.com/db2

IBM Information Management System: ibm.com/ims

IBM solidDB: ibm.com/software/data/soliddb

IBM InfoSphere Streams: ibm.com/software/data/infosphere/streams

Will your z/OS batch jobs complete or capsize?

Smart/RESTART lets your applications restart from near the point of failure — after abends, recompiles, even system IPLs. Your applications can run restartably, often without source changes!

Smart/RESTART guarantees that your program's sequential file and cursor position, working storage and VSAM updates stay in sync with changes to DB2, MQ, IMS and other RRS compliant resources. So you can restart fast with assured integrity.

Smart/RESTART is a robust, reliable and proven solution used by Global 2000 organizations worldwide to run their mission-critical z/OS batch applications. It's the standard for z/OS batch restart.

Restart made simple™

Download our White Paper:

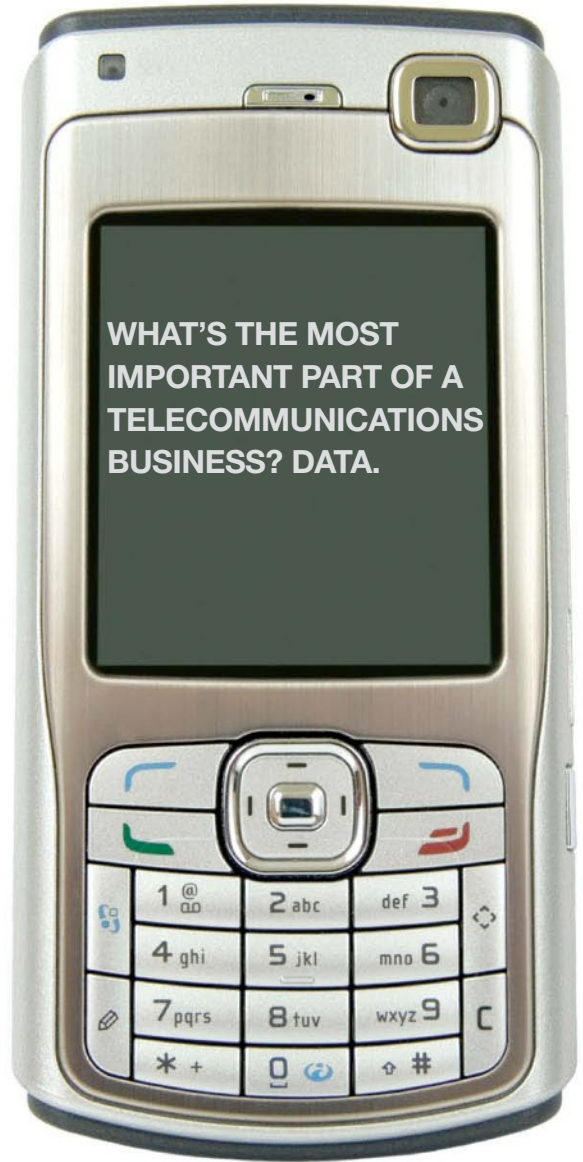
*"Beyond Restart and Concurrency:
z/OS System Extensions for
Restartable Batch Applications"*

For a free trial visit www.relarc.com, or call +1 201 420 - 0400

rai Relational
Architects
International

DB2, Websphere MQ and IMS are registered trademarks of IBM Corp.

WHAT'S THE MOST
IMPORTANT PART OF A
TELECOMMUNICATIONS
BUSINESS? DATA.



CALL ME

BY TAM HARBERT

D

ata is the lifeblood of telecommunications companies. Unless they keep the data flowing, they won't have a business. But to grow, or even sustain, that business, companies must do more with that data than ever before.

Telecommunications companies face enormous competitive pressures. Voice—the original telco application—has become a commodity. More and more households are dropping landlines altogether in favor of mobile. And providers at every level must contend with fickle, price-conscious consumers who switch providers seemingly on a whim. Consequently, telcos are desperately trying to offer new products, services, and bundles to retain current customers, attract new ones, and encourage all of them to consolidate their spending with one provider. Ultimately, operators want to become one-stop shops for all services and thereby increase their average revenue per user. “Every service provider is moving toward a converged offering,” says Arvind Sathi, lead architect in the communications sector of the IBM Software Group.

This shifting market creates huge data management challenges. Telcos rely on lightning-fast database applications to route and connect voice calls and data services. In a typical mobile call, for example, the system must locate and identify the subscriber, what type of plan the subscriber uses, how much time or money is in the account, and other information. The operator may also integrate and optimize third-party value-added services, such as Short Message Service (SMS).

During the call, the operator gathers more than 100 pieces of data that by regulation must be stored in a call detail record (CDR). Many providers are combining that data with other information—such as caller location—in real time to enable innovative new services. All the while, the companies are merging and leveraging the flood of data from new services to provide a better customer experience and increase selling opportunities.

An avalanche of customer data

At the most basic level, telecommunications companies must inventory, manage, and provision circuits across the network. Reliability is critical. “Some of our online systems handle 5 million to 10 million transactions a day,” says Paul Gandolfo, principal performance analyst at Telcordia, an independent service provider whose clients include top-tier telecommunications operators. Telcordia provides software and services, based on the IBM Information Management System (IMS) database management system, that help telcos manage the network by, for example, tracking physical and logical resources to maximize resource utilization, reduce

Qwest and InfoSphere

The company: Qwest Communications is a major provider of network, data, and voice services in the United States. For residential customers, the company offers Internet service, digital phone service, wireless service through a partnership with Verizon Wireless, and TV through a partnership with DIRECTV. Qwest also provides network, data, and voice services to Fortune 500 companies.

The challenge: Qwest, as have most major providers, has expanded to offer many different products and services. Its systems, however, were originally designed to support particular products rather than the range of services it offers today. The lack of integration of data from those services inhibits customer service and prevents Qwest from maximizing revenue.

“The systems that were built to support different product silos need to be aware of each other, and they need to share data,” says Sandeep Kulkarni, vice president of IT operational support system (OSS) development at Qwest. Information on the service availability was not available to the customer service agents at the right time. And even if it were, the agents might not have information about whether the customer had the right mix of services to new service, he explains. For example, “if you’re going to sell them an over-the-top video service, you have to know whether broadband service has enough speeds to support the video in their region,” he says. On the billing end, different systems were using different discounting engines, so sometimes a customer was quoted one price when ordering, only to receive a bill listing a different price.

The technology: Qwest is using IBM InfoSphere Master Data Management Server for Product Information Management to develop an enterprise product catalog (a catalog that is shared between ordering and billing), along with Selectica (a rules-based engine designed for product configuration) and Simple Order from IBM Customer Care. Although the system is not yet fully deployed, early metrics show that Qwest will be able to achieve a cost reduction of \$25 million, a significant reduction in calls to customer service, and increased up-sell and cross-sell activities, according to Kulkarni.

“The systems need to be aware of each other, and they need to share data.”

— Sandeep Kulkarni, Vice President of IT OSS Development, Qwest

errors, and resolve network performance problems. “Our customers bet their businesses that these systems are going to be there and work without problems and with high availability,” Gandolfo says.

Beyond the network operation itself, the operator manages data about customers and the services to which they subscribe, and most of that data must be accessed in real time. A mobile call involves “a highly and, at times, almost impossibly complex exchange of digital information,” says Charles King, president and principal analyst at Pund-IT, Inc.

The amount of such “call context data” has grown dramatically, says Ari Valtanen, director and CTO for solidDB in the IBM Software Group. IBM solidDB is a relational, in-memory database designed for such real-time, mission-critical applications. “Ten years ago, about one kilobyte was the norm,” says Valtanen. “Now it’s common to use tens to the low hundreds of kilobytes, and I’ve seen a few companies designing for as much as 1 megabyte per call or session. All that data has to be kept in memory and managed efficiently.” In addition to availability, high throughput and low latency are key in these applications. The entire call must be set up in 300 milliseconds, which means these database queries must be done in microseconds, says Valtanen.

Meanwhile, the system collects and stores the CDRs, each of which contains 140 fields of information, such as where the call originated and terminated, what rate applied, and the length of the call, according to Brian Kirk, vice president of business development at NetworkIP and its subsidiary, Jaduka.

Discovering new services and new customers

In their push to remain competitive and offer innovative services, operators are increasingly combining data in new ways. This information “provides a rich resource for data mining, developing new services, and entering new commercial markets, like targeted ad campaigns,” notes King. In Europe, for example, some network operators correlate the location of the caller with businesses in that area. They then use that information to send special offers, such as a coupon for the Starbucks on the corner, says Valtanen.

Effectively combining and leveraging this proliferation of data also has the potential to cut operating costs and improve customer satisfaction. Although providers are trying to converge their services, the data from each service has typically remained in a separate database. Even within

each service, the data from sales or from technical support might not be available to customer service call centers in a form that they can use. No company has managed to achieve such mega data management yet, but at least one—Qwest—is pursuing that path.

Another type of data that operators must handle is the information that customers themselves store on the network, such as text messages and e-mails. Database technology can enable more efficient storage of such information, says King. IBM DB2, for example, incorporates data compression features that enable telcos to store data on fewer arrays. This capability not only saves costs, but also enables quicker and more effective data backups, he says.

It all goes to show how important the flow of data is to telecommunications companies. Get this lifeblood flowing in the right direction, and the business stays healthy and vibrant. But if it's lost or trapped in clogged arteries, the business will at best be stagnant, at worst, on life support. *

Tam Harbert is a Washington, D.C.-based journalist who covers technology, business, and public policy.

Soprano Design and DB2

The company: IBM Business Partner Soprano Design provides a messaging infrastructure for Short Message Service (SMS) and Multimedia Messaging Service (MMS) messages. Mobile network operators, wireless application service providers, and others integrate the Soprano Design software into their networks and resell it to their customers. The platform enables companies to use messaging to better communicate with staff, customers, and suppliers.

The challenge: The application requires industry-standard technology that's easy to administer, manage, and integrate, so "we can spend less time worrying about how to manage and operate the database and focus more on enhancing and providing rich applications and staying ahead of the market," says Mohamed Odah, general manager of Soprano Design business operations in Australia.

The technology: Using IBM DB2 9.7—the latest version—and working with the IBM Express Runtime team, Soprano developed a DVD-installable image of its software that significantly eases installation for its customers. "It's a click of a button and everything is installed," says Odah. "We could not get that done with any other technology."

The DVD not only helps customers, but also saves Soprano time and resources. Previously, Soprano technicians spent four to five days installing the software at the customer site.

NetworkIP and Informix

The company: NetworkIP offers telephony service to the prepaid international calling market, primarily through calling cards that are rebranded by other telecom operators. Through its switching network and proprietary software, NetworkIP routes international calls while ensuring high-quality connections in more than 120 countries. The company also offers a Web-conferencing service to small and midsized businesses and, through its subsidiary Jaduka, telephony services to large enterprises.

The challenge: The company needed to increase capacity while ensuring high quality and reliability of calls. The company's software

manages the network and account information, determining the route of the call, the pricing, how much money is left on the calling card, the branding or marketing messages to be delivered through that account, what language to use, the menu options available, and the correct customer service number for that account. Each call generates 100 to 150 SQL statements, according to Andrew Ford, senior database administrator at the company. That amounts to 375 million database queries a day.

Even one database lock can affect users immediately. "It's one thing to be unable to run a report for a little while, but when you're making calls,

every single call depends on having access to that data," says Brian Kirk, vice president of business development at NetworkIP and Jaduka.

The technology: To increase speed and capacity, the company upgraded to IBM Informix 11.5, along with upgrades in server and storage hardware. With the

upgrade, NetworkIP increased its capability to manage accounts from 1.3 billion to 3 billion accounts, according to Ford. Benchmarks show that the system could now handle 5 million account transactions per hour, compared to 720,000 before the upgrade, he adds.

RESOURCES

IBM Information Management System: ibm.com/ims

IBM solidDB: ibm.com/software/data/soliddb

IBM Informix Dynamic Server: ibm.com/informix/ids

IBM InfoSphere Master Data Management Server for Product Information Management: ibm.com/software/data/infosphere/mdm_server_pim


Unleashing

the Value of Solid-State Drives

for DB2 Workloads

A little solid-state speed can go a long way—here's how to make the most of low-latency storage

By Sunil Kamath



Over the years, database administrators and database application developers have wrestled with storage layout and disk configuration when implementing mission-critical transactional, data warehouse, or mixed workloads. They spend tremendous amounts of time and money designing and optimizing applications for efficient I/O.

Now, the most commonly used enterprise hard disk drives (HDDs) are limited by the rate of head movement, the speed of the spinning platter, and seek latency. Solid-state drives (SSDs, also known as flash drives) address these challenges, providing fast, low-latency data access at reduced energy consumption (IOPS/watt); Figure 1 shows an approximate comparison of key SSD metrics relative to enterprise HDDs. Recently, the density of SSDs has tremendously improved; 200 GB and larger drives already are available, and capacity is expected to double during the next six months. These dramatic improvements have made SSDs prime candidates for use in performance-sensitive or mission-critical database applications.

In many cases, organizations will be able to place all of their databases on SSDs. But although SSD costs are declining every quarter, SSDs are still more expensive

than HDDs. Nevertheless, with IBM DB2, organizations can improve overall application performance by migrating carefully chosen data to higher-performance storage media. This article explores how a team at IBM tested the behavior of SSDs with DB2 under an online transaction processing (OLTP) workload, and how we found ways to determine which data to place on SSDs to derive the maximum benefit from limited SSD resources.

Taking advantage of SSD storage with DB2

The rich partitioning options offered by DB2 let you take advantage of different classes of storage. Specifically, you can use range partitioning to split large tables and indexes into smaller partitions organized by range (typically date range, but other parameters are possible), and place them on separate tablespaces backed by SSD storage. Then use the range partition features of DB2 to roll newer data into the frequently used or “hot” tablespaces as the older data grows colder.

To take advantage of low-latency SSD resources, you can also identify hot tables and then use the `ADMIN_TABLE_MOVE` stored procedure to move the table onto the tablespace backed with SSD storage.

Selecting hot data can be valuable to make the most of your SSD resources. SSDs are best suited for random I/Os, where they can improve performance over 15,000 rpm HDDs by more than 100x (measured in IOPS), compared to 10x improvements for sequential I/Os over HDDs.

To confirm placement options for DB2 objects on SSDs, we chose an IBM DB2 OLTP workload to drive a series of tests, starting with a baseline of all database files on HDDs. When we moved the entire database onto SSDs, we achieved a 10x performance improvement (measured in transactions executed per second). [Editor’s note: *Lab tests of IBM Informix Dynamic Server with SSDs also have revealed performance improvements.*] We then tested several scenarios after migrating portions of the database application to SSDs.

We started with a blind placement strategy of putting the database index files on SSD storage, but that generated only a modest improvement (less than 2x). It became clear that to drive greater performance improvements, we needed to reserve SSD space for the database objects that drove the highest amount of random physical I/O traffic and that were most critical to application performance. After some experimentation with DB2 monitoring data and operating system performance metrics, we achieved an optimal placement of tables and indexes on SSDs that delivered more than 8x performance improvement with only 30 percent of the overall database on SSDs (see Figure 2).

Approximate metrics as seen from application	15,000 rpm HDD	SSD
IOPS	150–300	1,000–20,000
Response time (for read)	5–7 ms	1–3 ms
I/O bandwidth	30–60 MB/sec	200–700 MB/sec
IOPS/GB	1–3	20–100
Power (watts)	4–6 watts	5–8 watts

Figure 1: SSDs easily outpace the 15,000 rpm HDDs commonly used for enterprise applications

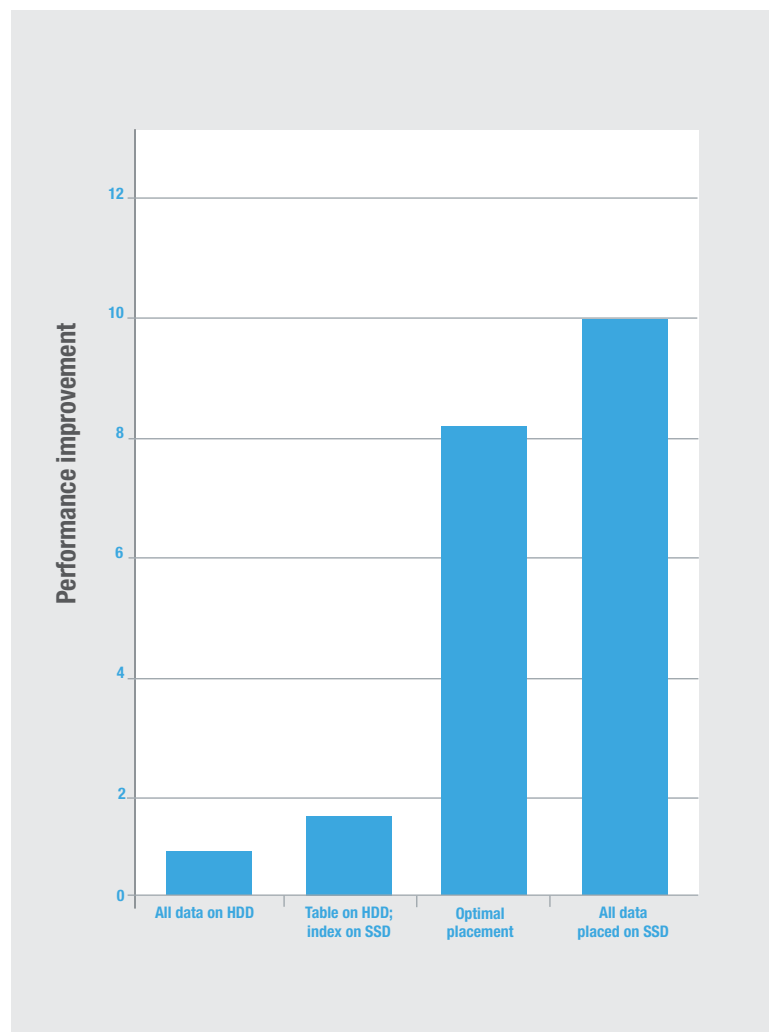


Figure 2: Identifying high-use data can help maximize the performance benefits of SSDs

Cold, warm, and hot: Identifying in-demand data

So, what's the best way to determine which DB2 tables or indexes to place on SSDs? Prior application knowledge based on the data model can be a good place to start. But although this type of intuitive tuning is valuable, we wanted to develop a more objective way to evaluate which files and data would provide the most performance payback if they were placed on SSD. Constantly determining which data is hot and cold and then moving that data across storage tiers may additionally burden a DBA.

You can use operating system monitoring tools, such as AIX filemon and iostat utilities, to determine which logical volumes, file systems, and/or disk volumes are hot and therefore driving more IOPS and higher access latency.

DB2 provides snapshot monitoring capabilities that can be used for performance monitoring at various levels for all connected applications—database, buffer pool, tablespace, table, application, lock, and others. This information can be collected via DB2 command line processor (CLP) using snapshot commands, SQL table functions, or by using snapshot monitoring application programming interfaces (APIs) via C/C++ programs. Additionally, the IBM DB2 Performance Expert tool with Extended Insight feature can also be used to obtain rich insight based on current and historical monitoring data. By analyzing these monitoring counters, you can get valuable insight into the I/O characteristics of database applications.

By combining data from filemon, iostat, and DB2 snapshot monitoring with a buffer pool monitoring switch, you can develop the following information:

- ▶ Tablespace page size (4 KB, 8 KB, 16 KB, or 32 KB)
- ▶ Number of used pages (defined in units of tablespace page size)
- ▶ Buffer pool data physical reads (defined in units of tablespace page size)
- ▶ Buffer pool index physical reads (defined in units of tablespace page size)
- ▶ Buffer pool xda physical reads (defined in units of tablespace page size)
- ▶ Buffer pool read time (in seconds)

From this data, you use the equations in Figure 3 to calculate four metrics that help determine which tablespaces are better candidates to place on SSDs:

- ▶ **Access density:** The number of physical I/Os relative to the number of used pages in the tablespace
- ▶ **Access latency:** A measure of latency for those physical I/Os
- ▶ **Relative tablespace weight:** A function of access density and access latency
- ▶ **Sequentiality ratio:** A measure of how random or sequential the I/O accesses to a given tablespace are

TABLESPACE METRICS

$$\text{Total physical I/Os} = \frac{(\text{Buffer pool data physical reads} + \text{Buffer pool index physical reads} + \text{Buffer pool xda physical reads} + \text{Buffer pool temporary data physical reads} + (\text{Direct reads} * 512))}{(\text{Tablespace page size})}$$

$$\text{Page velocity} = \frac{(\text{Total physical I/Os})}{(\text{Snapshot interval in seconds})}$$

$$\text{Access time} = (\text{Total buffer pool read time} + \text{Direct reads elapsed time})$$

$$\text{Access density} = \frac{(\text{Page velocity})}{(\text{Number of used pages in tablespace})}$$

$$\text{Access latency} = \frac{(\text{Access time})}{(\text{Total physical I/Os})}$$

$$\text{Weighting factor} = (\text{Access density}) * (\text{Access latency})$$

$$\text{Sequentiality ratio} = \frac{(\text{Asynchronous pool data page reads} + \text{Asynchronous pool index page reads} + \text{Asynchronous pool xda page reads})}{(\text{Buffer pool data physical reads} + \text{Buffer pool index physical reads} + \text{Buffer pool xda physical reads})}$$

Figure 3: The weighting factor and sequentiality ratio metrics will help determine which tablespaces to move to SSDs

“So far, our focus has been making SSD technology available through our hardware. Today, we provide the tools for our customers to identify and move hot data to SSD. In the future, IBM systems will take those actions for them.”

—Andy Walls, *Distinguished Engineer, Storage Hardware Architecture, IBM*

When these metrics are summarized for all tablespaces based on descending order of weighting factor, those tablespaces that have a higher weighting factor are better candidates for SSDs. Tablespaces that have a lower sequentiality ratio also are better candidates for SSDs.

More performance options

The weighting factor and sequentiality ratios calculated in Figure 3 will help you find the tablespaces that are the best candidates for placement on SSDs. However, depending on your environment and available SSD resources, there are two more options that you should be aware of.

First, consider placing DB2 temporary tablespaces on SSDs if you have the capacity. This can help improve the sort time driven by complex “group by” or “order by” queries. This is especially useful for data warehouse workloads that usually drive a large number of sorts and therefore spill into disk.

Finally, active log files are less desirable candidates for placement on SSDs, as these operations usually drive sequential I/Os and are typically hit on storage write cache. However, if the application is suffering from large commit times and increased log I/O latency, placing the active logs on SSDs can help. *

Sunil Kamath is a senior technical staff member and senior manager in Information Management at the IBM Toronto Software Labs.

Putting solid- state storage to work

It will be some time before an enterprise will be able to simply replace its HDDs with SSDs. However, IBM is integrating SSDs as part of its server and storage lines, and developing technologies that take advantage of SSD performance.

IBM has announced the availability of SSDs for use within the IBM System Storage DS8000 and DS5000. IBM has also added SSD-awareness to the automated data placement facility on IBM System z. The Data Facility—System Managed Store (DFSMS) management software allows users to define storage pools and data classes, and the DFSMS facility then automatically provisions the data requiring the highest performance to the fastest storage.

SSD technology can also be used as part of a storage area network (SAN) through the IBM SAN Storage Volume Controller (SVC), which now contains software based on the Project Quicksilver proof of concept that provided logical unit number (LUN) management and SAN access to SSD data with over 1 million IOPS.

Looking ahead, the next phase of SSD integration will be systems that identify critical workloads or “hot” data and then automatically migrate that data to low-latency storage resources like SSDs. “So far, our focus has been making SSD technology available through our hardware,” says Andy Walls, distinguished engineer, Storage Hardware Architecture at IBM. “Today, we provide the tools for our customers to identify and move hot data to SSD. In the future, IBM systems will take those actions for them.”

RESOURCES

IBM solid-state storage: ibm.com/systems/storage/solutions/ssd/

IBM DB2 for Linux, UNIX, and Windows:
ibm.com/software/data/db2/linux-unix-windows/

IBM DB2 9.7 information center:
<http://publib.boulder.ibm.com/infocenter/db2luw/v9r7/index.jsp>



solidDB and the **Secrets** of **Speed**

*By Antoni Wolski and
Sally Hartnell*

How the IBM in-memory database redefines high performance

A RELATIONAL IN-MEMORY DATABASE, IBM SOLIDDB IS USED worldwide for its ability to deliver extreme speed and extreme availability. As the name implies, an in-memory database resides entirely in main memory rather than on disk, making data access an order of magnitude faster than with conventional, disk-based databases. Part of that leap is due to the fact that RAM simply provides faster data access than hard disk drives.

But solidDB also has data structures and access methods specifically designed for storing, searching, and processing data in main memory. As a result, it outperforms ordinary disk-based databases even when the latter have data fully cached in memory. Some databases deliver low latency but cannot handle large numbers of transactions or concurrent sessions. IBM solidDB provides throughput measured in the range of tens-to-hundreds of thousands of transactions per second while consistently achieving response times (or latency) measured in microseconds.

This article explores the structural differences between in-memory and disk-based databases, and how solidDB works to deliver extreme speed.

Some RDBMS history

When the first data management systems emerged in the 1960s, disk drives were the only place to store and access large amounts of data in a reasonable time. RDBMS designers concentrated on optimizing I/O and tried to align the data access patterns with the block structure imposed by the drives. Design strategy frequently centered on a shared buffer pool where data blocks were kept for reuse, while advances in access methods produced solutions like the renowned B+ tree, which is a block-optimized index.

Meanwhile, query optimization tactics focused on minimizing page fetches wherever possible. In the fierce battle for performance, disk I/O was often the deadliest enemy, and processing efficiency was sacrificed to avoid disk access. For example, with typical page sizes of 8 KB or 16 KB, in-page processing is inherently sequential and less CPU-efficient than random data access. Nevertheless, it remains a popular way to reduce disk access.

When the era of abundant memory arrived, many DBAs increased their buffer pools until they were large enough to contain an entire database—thus creating the concept of a

fully cached database. But within the RAM buffer pools, the DBMSs were still hostage to all the structural inefficiencies of the block-oriented I/O strategy that had been created to deal with hard disk drives.

Moving past the blocks

One of the most noticeable differences of an in-memory database system is the absence of large data block structures. IBM solidDB eliminates them. Table rows and index nodes are stored independently in memory, so that data can be added without reorganizing big block structures.

In-memory databases also forgo the use of large-block indexes, sometimes called bushy trees, in favor of slim structures where the number of index levels is increased and the index node size is kept to a minimum to avoid costly in-node processing. The most common in-memory database index strategy is called T-tree. IBM solidDB instead uses an index called trie (or prefix tree), which was originally created for text searching but turns out to be perfect for in-memory indexing. A trie (the name comes from the word *retrieval*) is made up of a series of nodes where the descendants of a given node have the same prefix of the string associated with that node. For example, if the word “dog” were stored in a trie as a node, it would descend from the node containing “do,” which would descend from the node containing “d.”

Trie indexes increase performance by reducing the need for key value comparisons and practically eliminating in-node processing. The index contains a node that is a small array of pointers to the lower level. Instead of using the whole key value to walk the tree by way of comparisons, the key value is cut into small chunks of a few bits. Each chunk is a direct index to the pointer array of the corresponding level: the first left-hand-side chunk to the first-level nodes, the second chunk to the nodes of the second level, and so on. Thus, the entire search can be performed with just a few array element retrievals. Also, each index node is a small data block (approximately 256 bytes in solidDB), which is beneficial because the blocks fit precisely into modern processor caches, increasing processing efficiency by promoting efficient cache use. Small data arrays like these are the most efficient data structure in modern processors, and solidDB uses them frequently to maximize performance.

Checkpoints and durability: Paths to speed

IBM solidDB uses several additional techniques to accelerate database processing, starting with a patented checkpointing method that produces a snapshot-consistent checkpoint without blocking normal transaction processing. A snapshot-consistent checkpoint allows the database to restart from a checkpoint only. Other database products do not normally allow that—the transaction log files must be used to recalculate

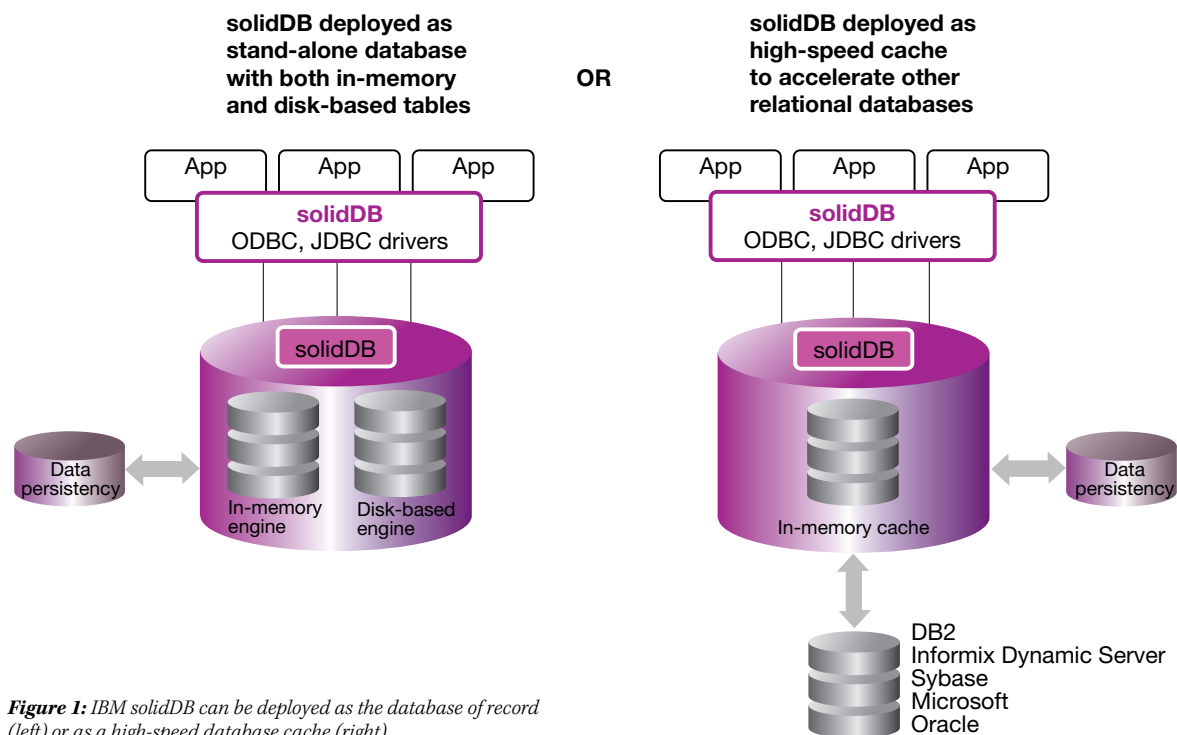


Figure 1: IBM solidDB can be deployed as the database of record (left) or as a high-speed database cache (right)

Dressed to the nines

Almost anyone who has spent any time around highly available systems has heard the term “five nines” used to refer to uptime. It’s a shorthand way to say that a system is available 99.999 percent of the time. So if the subsecond failover capabilities of solidDB make it possible to achieve six nines of availability, it’s easy to imagine just adding another nine to that percentage: 99.9999.

But unless you’ve done the math, it’s difficult to appreciate the real-world impact of reliability statistics. The following table should help put things in perspective.

Availability	Approximate downtime per year
95%	18 days, 6 hours
99%	3 days, 11 hours, 18 minutes
99.9%	8 hours, 20 minutes
99.99%	50 minutes
99.999%	5 minutes
99.9999%	30 seconds

the consistent state (solidDB allows transaction logging to be turned off, if desired). The solidDB solution is made possible by the ability to allocate row images and row shadow images (different versions of the same row) without using inefficient block structures. Only those images that correspond to a consistent snapshot are written to the checkpoint file, and the row shadows allow the currently executing transactions to run unrestricted during checkpointing.

Further, the solidDB query optimizer recognizes the different nature of the in-memory tables by estimating execution costs in a new way. Query optimization focuses on CPU-bound execution paths, while a fully cached database will still be preoccupied with optimizing page fetches to mass storage that are no longer an issue.

Another technique IBM solidDB uses is the relaxation of transaction durability. In the past, databases always supported full durability, guaranteeing that the written data is made persistent the moment the transaction is committed. The problem is that full durability inflicts synchronous log writes, and thus it consumes resources and reduces response times. In many situations, accepting less durability for some tasks

for the sake of faster response times is a perfectly acceptable trade-off. With solidDB, transaction durability can be relaxed at run time for a given database session or even for a single transaction.

IBM solidDB also increases database performance by helping developers avoid process context switches in client/server interactions. By using a database access driver provided with solidDB that contains the full query execution code, a developer can effectively link the application with the DBMS code and use shared memory to share the data among the applications.

When all of these measures are applied and the application load is of a type that would inflict significant I/O in a traditional database, the increased throughput using solidDB can be an order of magnitude higher. Further, response time improvements are even more dramatic: latencies for query transactions are usually 10 to 20 microseconds and latencies for update transactions are generally less than 100 microseconds. In a traditional disk-based database, the corresponding times are typically measured in milliseconds.

solidDB speed and power

Beyond these performance advantages, solidDB also provides several additional benefits. It combines a fully transactional in-memory database and a powerful, disk-based database into a single, compact solution with the ability to transparently host part of the same database in memory and part on disk. And IBM solidDB is the only product on the market that can be deployed as a high-speed cache in front of almost any other relational, disk-based database (see Figure 1). Finally, solidDB delivers extreme availability, going beyond the typical five nines to 99.9999 percent uptime (see sidebar, “Dressed to the nines”). In other words—if you’re looking for extreme speed, you’ll find it, but that’s just the beginning for IBM solidDB. ✱

Antoni Wolski is chief researcher for solidDB. With more than 20 years of activity in database technology, he has contributed to solutions, products, research, patents, and literary work in the field.

Sally Hartnell joined IBM from the acquisition of Solid Information Technology, and she remains responsible for worldwide marketing of IBM solidDB.

RESOURCES

IBM solidDB: ibm.com/software/data/soliddb/

IBM solidDB checkpointing method: www.computer.org/portal/web/csdl/doi/10.1109/ICDE.2006.140

What Is DB2 pureScale?

Going to extremes on scale and availability for DB2

By Chris Eaton and Paul C. Zikopoulos

IN OCTOBER 2009, IBM ANNOUNCED A NEW TECHNOLOGY, primarily aimed at online transaction processing (OLTP) scale-out clusters, called IBM DB2 pureScale. DB2 pureScale is a new feature that provides scale-out active-active services for IBM DB2 running on AIX on Power Systems servers. It's designed to deliver the highest levels of distributed availability and scalability wrapped in a well-thought-out, up-and-running path that's much easier to operate than other clustered database systems. In this article, we'll give you the basics of DB2 pureScale from a technology perspective, showing how DB2 pureScale delivers both transparent application scalability and extreme availability.

I see what you did there

If you're familiar with data sharing on DB2 for z/OS, then the DB2 pureScale architecture may look very similar—that's because it is! IBM took the fundamental tenets of DB2 for z/OS data sharing and coupled them with the most current distributed technologies to deliver unprecedented availability and scalability services to distributed platforms. We'd like to note one thing here: DB2 running on System z servers already delivers first-rate availability. For example, Toronto Dominion Bank (TD Bank) has had 100 percent availability of customer information for 10 consecutive years, including two DB2 for z/OS upgrades during that timeframe. Even the CEO of our biggest competitor said of DB2 for z/OS: "It's a first-rate piece of technology."¹

Figure 1 shows an example of a DB2 pureScale environment. A DB2 server that belongs to a pureScale cluster is called a member; each member can simultaneously access the same database for both read and write operations. Currently, the maximum number of members in a pureScale cluster is 128.

The IBM PowerHA pureScale server provides centralized lock management services, a centralized global cache for data pages (known as the group buffer pool), and more. Each member in a DB2 pureScale data-sharing group can interact directly with the PowerHA pureScale server through an InfiniBand network using User Direct Access Programming Library (uDAPL), a non-messaging base protocol that provides each member with point-to-point connectivity to the centralized locking and caching services.

Local agents, cluster-wide reach

Transparent application scaling means that applications don't have to be cluster-aware to truly take advantage of the scale-out architecture. To deliver this scaling, DB2 pureScale uses remote direct memory access (RDMA) technology along with PowerHA pureScale technology to eliminate communication between members for lock management and global caching services.

¹ Matthew Symonds, "In Larrys Own Words," *eWeek*, October 31, 2003, <http://www.eweek.com/c/a/Database/In-Larrys-Own-Words/2/>.

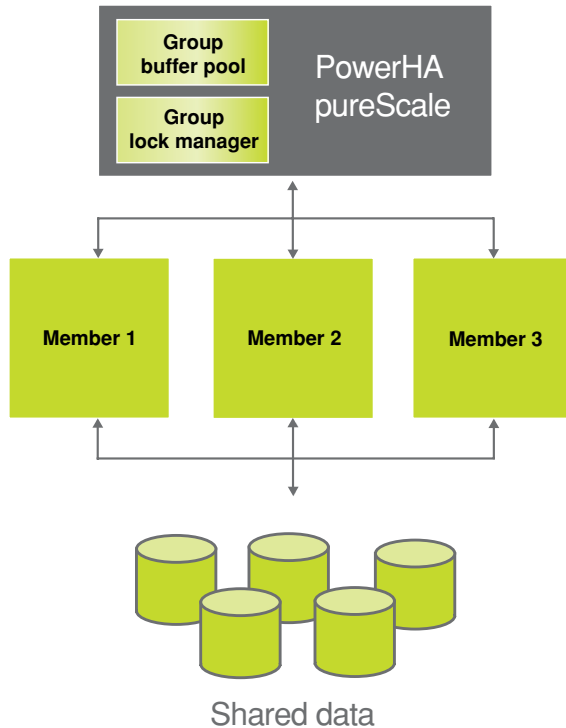


Figure 1: In a DB2 pureScale cluster, each member has direct memory-based access to the centralized locking and caching services of the PowerHA pureScale server

RDMA allows each member in the cluster to directly access memory in the PowerHA pureScale server, and vice versa, in microseconds. For example, assume that Member 1 in Figure 1 wants to read a data page that isn't in its local buffer pool. DB2 assigns an agent (or thread) to perform this transaction. The agent then uses RDMA to directly write into the memory of the PowerHA pureScale server to indicate that it has interest in a given page (this is called a read-and-register request). If the page that Member 1 wants to read is already in the global centralized buffer pool, the PowerHA pureScale server will push that page directly into Member 1's memory instead of having the agent on that member perform the I/O operation to read it from disk. Effectively, RDMA allows a member's agent to simply perform what appears to be a local memory copy operation, when in fact the target is the memory address of a remote machine.

These lightweight remote memory calls, along with a centralized buffer pool and lock management facilities, mean that an application does not have to connect to the member where the data already resides to achieve scalability. It is just as efficient for any member in the cluster to receive a data page from the global buffer pool, regardless of the size of the cluster. Most RDMA calls are so fast that the DB2 agent making the call doesn't even need to yield the CPU while

waiting for the response. For example, to notify the PowerHA pureScale server that a row is about to be updated (and therefore an X lock is required), a member's agent performs a Set Lock State (SLS) request by writing the lock information directly into the PowerHA pureScale server's memory. The entire round-trip for this SLS operation can take less than 15 microseconds and therefore the agent likely doesn't need to yield the CPU.

Does your cluster know where your pages are?

DB2 pureScale takes availability to a whole new level. If a member in a DB2 pureScale cluster fails, DB2 provides full access to every page of data that doesn't need recovery. What's more, without performing a single I/O operation, DB2 is aware at all times of the specific pages that are in need of recovery.

How does this happen? Every time a member reads a page into its buffer pool, the PowerHA pureScale server not only keeps track of this "interest," but also requests from members to update rows on those pages. Whenever an application commits a transaction, dirty pages are written directly into the PowerHA pureScale server. If any member fails, the PowerHA pureScale server has a list of pages that the failed member was in the process of updating as well as the pages that were updated and committed by the failed member but weren't yet written to disk.

When a failure occurs on a shared disk cluster, it's critical that no other node in the cluster reads or updates from disk any pages that might not have been recovered yet. Because the PowerHA pureScale server knows which pages were in the process of being updated by the failed node, and the PowerHA pureScale server already has the dirty committed pages from that member in its centralized buffer pool, DB2 pureScale doesn't need to block other members from continuing to process transactions while it locks the pages that need recovery.

What's more, the act of recovery in DB2 pureScale happens very quickly. Each member has processes that are sitting idle but are ready if a failure occurs. Should a member fail, one of these recovery processes is activated; since these processes already exist, there's no need for the operating system to waste valuable time to create a process, allocate memory to it, and so on. This recovery process instantly begins to prefetch dirty pages from the centralized buffer pool into its own local buffer pool. In the majority of cases, this recovery won't require additional I/O operations because the pages that need recovery probably are already in the centralized buffer pool and can be transferred in microseconds using RDMA. Meanwhile, all other applications on all other members continue to process transactions on any page that doesn't need recovery and read pages from disk because the PowerHA pureScale server knows which pages on disk are clean and which need recovery.

For typical transactional workloads, the time from the member failure until the time the pages are recovered and available to another transaction is typically 20 seconds or less. Note that this recovery time includes the failure detection times, which many vendors exclude when referring to recovery times.

Finally, it is worth noting that components in the cluster—including the PowerHA pureScale server itself—are redundant. DB2 pureScale allows duplexing of the PowerHA pureScale server capability, so that locking information and shared cache information are stored in two separate locations in case the primary fails.

Summary

We've only scratched the surface of pureScale here. A lot of engineering is going on behind the scenes, so be sure to check out the links in the Resources sidebar. The bottom line is that for applications that need the highest levels of availability in a scale-out active-active configuration, DB2 pureScale delivers leading-edge capabilities to enhance your business continuity. And, with transparent application scalability, you no longer need to build cluster-aware applications in order to scale out to larger numbers of servers. ✱

Chris Eaton is technical evangelist and senior product manager for DB2, primarily focused on planning and strategy. He is an award-winning speaker, having spoken at worldwide DB2 conferences, and he has authored many DB2 books. Eaton also has one of the most popular blogs about DB2 on the Web at <http://it.toolbox.com/blogs/db2luw>.

Paul C. Zikopoulos, B.A., M.B.A., is the program director for the DB2 Evangelist team at IBM. He is an award-winning writer and speaker with more than 15 years of experience with DB2. Zikopoulos has written more than 300 magazine articles and 13 books on DB2. You can reach him at: paulz_ibm@msn.com.

RESOURCES

IBM DB2 for Linux, UNIX, and Windows: ibm.com/db2/9

IBM DB2 pureScale home page: ibm.com/db2/9/editions-features-purescale.html

IBM DB2 pureScale business case: https://www14.software.ibm.com/webapp/iwm/web/signup.do?lang=en_US&source=sw-infomgt&S_PKG=db2-purescale-wp

IBM DB2 pureScale overview: <http://download.boulder.ibm.com/ibmdl/pub/software/data/sw-library/db2/brochures/db2-pure-scale.pdf>

Solving the DB2 Management Puzzle is Easy



Quest Management Suite Controls the Chaos of DB2 Administration

Managing your complex DB2 environment just became easier, thanks to Quest® Management Suite for DB2 LUW. It combines all of the functionality you need for DB2 on Linux, Unix, and Windows, so you can manage multiple aspects of DB2 administration and performance with a single solution. Whether it's administration, SQL tuning, workload analysis and trending or real-time diagnostics, you've got the power you need to quickly and easily manage and control your day-to-day DB2 activities.

Watch the pieces of DB2 management fall into place with Quest.

Read our new technical brief, "Managing DB2 with Toad® – A Guide for Oracle Pros" at www.quest.com/DB2control

QUEST SOFTWARE®
Smart Systems Management



Intel and IBM Collaborate to Double In-Memory Database Performance

“To meet the continually increasing demand for faster, more powerful information technology, IT managers must consider deploying high-performance computing platforms, like the Intel Xeon processor 5500 series, and optimized software, such as in-memory database technology, in their data centers. IBM is pleased to work with Intel to find solutions to today’s biggest information management infrastructure challenges, as our leadership benchmark results demonstrate.”

—Ari Valtanen
Director & CTO, IBM solidDB



With the rise of more complex multi-core, multi-threaded infrastructure, IBM and Intel have introduced new processing, memory, and database innovations that can be combined to significantly accelerate the transformation of volumes of data into useful business insights.

By coupling the IBM® solidDB® 6.3 in-memory database with servers based on the Intel® Xeon® processor 5500 series, enterprises can achieve double the performance of previous-generation installations.¹ This new combination can unlock even more value from customer data management investments and can begin to transform static data repositories into dynamic solutions that provide information on demand.

IBM solidDB 6.3: Accelerates Information On Demand

IBM solidDB is relational, in-memory database technology that delivers extreme speed, performing up to 10 times faster than conventional, disk-based databases. An in-memory database exists to meet the performance demands of real-time applications requiring extreme speed and predictable response times. As the name implies, an in-memory database resides entirely in main memory rather than on disk, making data access extremely fast. There are more than 3 million deployments of solidDB in telecommunications networks, enterprise applications, and embedded software and systems. Market leaders such as Cisco, HP, Alcatel, and Nokia Siemens rely on solidDB for their mission-critical applications for reasons such as:

Extreme speed: Data resides in main memory at all times rather than on disk, enabling hundreds of thousands of transactions per second.

Extreme availability: Two copies of the data are synchronized at all times, allowing sub-second failover capability.

Low cost: Technology that is easy to deploy, administer, and embed directly into applications runs virtually unattended for lower total cost of ownership.

Intel Xeon Processor 5500 Series: A New Generation of Intelligent Servers

Enabling the newest generation of high-performance and energy-efficient computing, the Intel Xeon processor 5500 series can automatically adjust server performance and power consumption, or allow manual IT control to meet unique service level requirements. The new Intel Xeon processor 5500 series delivers up to 9x performance per server over single-core servers, enabling 9:1 server consolidation, up to 90 percent lower operating costs, and an estimated 8-month return on investment.²

Factors to consider include:

Intelligent performance: Intel® Turbo Boost Technology increases processor core speeds for more performance when workload conditions demand it.

Energy efficiency: Intel® Intelligent Power Technology lowers energy costs by automatically switching the processor and memory into the lowest available power state without sacrificing workload requirements.

High throughput: Intel® QuickPath Technology significantly lowers system latency and increases transaction processing bandwidth.

Innovations in Memory: The Key to Performance Leadership

The quad-core, eight-thread Intel Xeon processor 5500 series was built for throughput computing leadership. To help maximize application performance, Intel has introduced a completely new memory subsystem with three key innovations to reduce system latency while maximizing processing bandwidth. An advanced, three-tier cache architecture improves on-chip communications efficiency across processor cores and threads. A native DDR3 integrated memory controller significantly accelerates off-chip communications to main memory. Last, Intel QuickPath Technology introduces a dedicated point-to-point link to main memory and other processors that can deliver over twice the maximum transactional

throughput of previous-generation memory bus architectures.

IBM solidDB works under the premise that all data must be accessible with extreme speed, offering two performance advantages over conventional databases. First, because any data requested by an application is already in main memory, the need to transfer data blocks from disk to main memory is eliminated, significantly reducing application response times. Second, because of its optimized data structures and main memory access methods, solidDB transacts in memory significantly faster than disk-based databases, even if those databases cache all their data in main memory.

Figure 1 demonstrates that for existing customers with solidDB 6.1 deployed on previous-generation Intel Xeon processor-based servers, up to 2.75x more performance

can be realized from upgrading to solidDB 6.3 on servers based on the new Intel Xeon processor 5500 series.¹ And for customers who have already deployed solidDB 6.3 on previous-generation Intel processor-based hardware, up to 1.87x more performance is available from the Intel Xeon processor 5500 series on the same software configuration.

Learn More

Enterprises seeking to optimize the performance of their information management solutions must look at the underlying components in a new way. New processing, memory, and database innovations can be leveraged to significantly accelerate the transformation of volumes of data into useful business insights while saving energy costs and meeting customer service level demands.

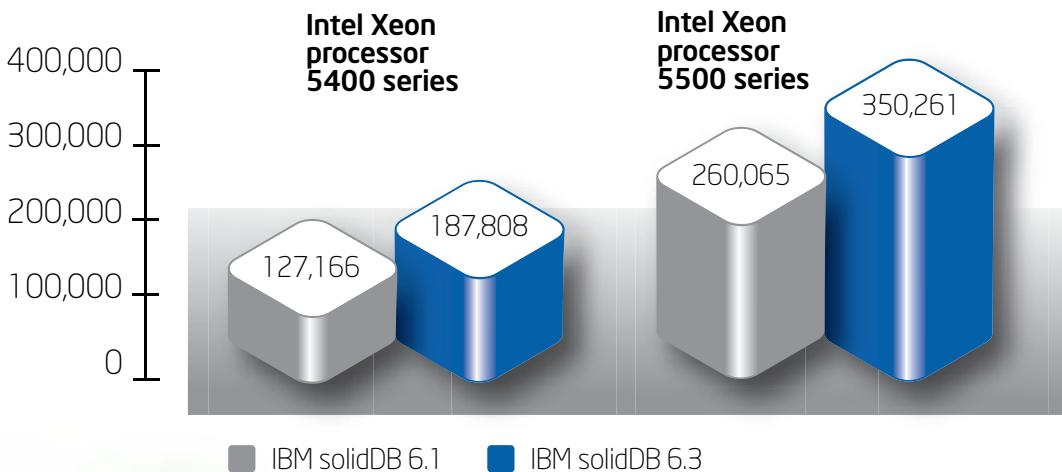


Figure 1: Transactional performance of IBM® solidDB® 6.3 on the Intel® Xeon® processor 5500 series scales 1.87x more than on previous-generation hardware systems, and 2.75x more than previous-generation hardware and software configurations.

¹ Source: IBM internal measurements as of March 2009. All of the benchmark results presented are derived using the TATP benchmark simulating a HLR application workload including a database populated with 1 million subscribers. The resulting database size was approximately 1.5 GB. In each case, the transaction mix was set to 80 percent read accesses and 20 percent write accesses. Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments might vary significantly. Users of this document should verify the applicable data for their specific environment.

² Source: 8-month ROI claim estimated based on comparison between 2S Single Core Intel® Xeon® 3.80 with 2M L2 Cache and 2S Intel® Xeon® X5570-based servers. Calculation includes analysis based on performance, power, cooling, electricity rates, operating system annual license costs, and estimated server costs. This assumes 8kW racks, \$0.10 per kWh, cooling costs are 2x the server power consumption costs, operating system license cost of \$900/year per server, per-server cost of \$6,900 based on estimated list prices and estimated server utilization rates. All dollar figures are approximate. Performance and power comparisons are based on measured SPECjbb2005* benchmark results (Intel Corporation, February 2009). Platform power was measured during the steady-state window of the benchmark run and at idle. Performance gain compared to baseline was 9x while the platform power was 0.8x.

-Baseline platform: Intel server platform with two 64-bit Intel Xeon processor 3.80GHz with 2 MB L2 cache, 800 FSB, 8x1GB DDR2-400 memory, one hard drive, one power supply, Microsoft® Windows® Server 2003 Ent. SP1, BEA® JRockit® build P27.4.0-windows-x86_64 run with two JVM instances

-New platform: Intel server platform with two quad-core Intel Xeon processor X5570, 2.93 GHz, 8 MB L3 cache, 6.4QPI, 12 GB memory (6x2GB DDR3-1333), one hard drive, one power supply, Microsoft Windows Server 2008 Ent. SP1, BEA JRockit build P27.4.0-windows-x86_64 run with two JVM instances

Intel, the Intel logo, and Xeon are registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries. IBM, the IBM logo, and solidDB are trademarks of International Business Machines Corporation in the United States, other countries, or both. *Other names and brands may be claimed as the property of others.

Making Humongous Doable

Need to manage a gigantic database?
Let DB2 do the heavy lifting.



Robert Catterall
(rcatterall@catterallconsulting.com) is president of Catterall Consulting, a provider of DB2 consulting and training services.

Really “big” in the context of a data-serving system, used to mean a few hundred transactions per second, a terabyte or more of data, and a thousand or more database objects (tables and indexes). Today, a system with these characteristics would be thought of as “large,” but the bar for “huge” has been raised considerably: more than a thousand transactions per second, multiple terabytes of data, and tens of thousands of database objects.

Organizations around the world must deal with such huge systems in a cost-effective manner. The IBM DB2 development group has consistently delivered features and functions that address this need for ever-more efficient support of ever-larger databases. You could write a book on all that stuff, but I have about 1,200 more words at my disposal, so I’ll focus on two of my favorites: a new and game-changing scale-out solution for DB2 on the AIX/Power platform, and a DB2 for z/OS feature that is great for large databases but overlooked by a lot of DB2 people.

DB2 pureScale: Advanced shared-data clustering

You can read the specifics on DB2 pureScale in the article “What is DB2 pureScale?” in this issue. Here, I want to talk about why it’s so important. Vertical scalability, also known as scale-up, is a DB2 strength. Give a DB2 server more and/or faster engines, and you’ll get better throughput. Sounds simple, but making good on that proposition—once you get past a few processors—requires advanced engineering.

That said, sometimes a really big database system is best supported with a scale-out (multi-node) configuration. In such cases, it’s important to go with a scale-out solution that provides a good match for the requirements of the target application. For large data warehouse systems, the shared-nothing multi-node architecture implemented via the DB2 for Linux, UNIX, and Windows (LUW) data partitioning feature (DPF) makes lots of sense. For online transaction processing (OLTP) applications, on the other hand, a shared-data cluster (multiple servers with concurrent read/write access to a database on shared disk) is likely to be a better fit. For years, the only available DB2-based shared-data system was DB2 for z/OS data sharing on a mainframe parallel sysplex. That changed with the announcement of IBM DB2 pureScale, a shared-data solution for DB2 on the IBM AIX/Power platform.

DB2 pureScale, announced in October 2009, trumps the shared-data cluster competition in the UNIX market. Here’s the deal: if you’re going to give multiple data servers read/write access to one database, you have a couple of choices when it comes to keeping the different servers from trashing the consistency of said database. One option is to have a node directly communicate with all the other nodes regarding data rows that it’s changing and data that it has cached locally. Alternatively, you can go with a centralized mechanism by which a data server node posts global lock and global buffer pool information to structures residing in devices that provide a shared-memory resource to

Join us online

- ▶ **Quick content searches** throughout the entire issue
- ▶ **Direct links** to related community resources
- ▶ **Easy information access**, sharing, and printing



IBM
data
management
KNOWLEDGE. PERFORMANCE. RESULTS.

Visit ibm.com/dmmagazine and sign up for your complimentary digital subscription.

You'll get the same in-depth technical content, practical advice, and hands-on tips about how to improve productivity and performance as the print edition—now including real-world commentary about how your peers are using data and information architectures to reduce costs and improve business results.

IBM.COM/DMMAGAZINE

the group. The former approach works, but it doesn't scale very well. As the number of nodes increases, so does the amount of time that they spend telling each other what they're doing.

DB2 pureScale implements the centralized approach to global lock and data cache coherency. This is the same design that for years has delivered previously unattainable levels of scalability and availability to organizations using DB2 for z/OS in data-sharing mode. In essence, DB2 pureScale servers function as the coupling facilities in a DB2 for z/OS data-sharing group. They house the global lock, group buffer pool, and shared communications area structures in a super-high-performance shared-memory resource.

As DB2 for z/OS data sharing took the already-high scalability and availability standards of the mainframe DB2 platform and raised them still higher, so will pureScale for DB2 on AIX/Power—the platform that

DB2 pureScale really is a whole new ball game with respect to high-end, UNIX-based, OLTP-serving database systems, and it's going to be fun watching this one play out.

already sets the standard for UNIX system reliability. Other positive parallels between DB2 for z/OS data sharing and DB2 pureScale are worth noting: both are application transparent, provide system-managed workload balancing across nodes, and allow very granular increases in system processing capacity. It really is a whole new ball game with respect to high-end, UNIX-based, OLTP-serving database systems, and it's going to be fun watching this one play out.

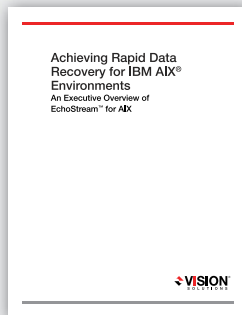
Letting DB2 for z/OS manage disk space allocation for you

From the much-trumpeted pureScale technology, we turn now to a DB2 for z/OS

capability, introduced with DB2 8, that a DB2 consultant friend of mine called “a real success story that never seems to get much press.” He was referring to automatic management of primary and secondary disk space allocation for DB2-managed (that is, **STOGROUP**-defined) data sets. If you have a database with lots of tablespaces and indexes, this is indeed a sweet feature.

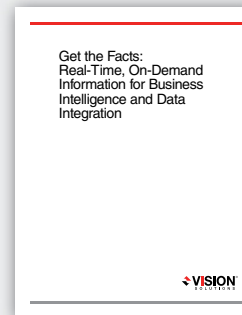
Many DB2 for z/OS DBAs have spent lots of time determining appropriate primary and (especially) secondary space-allocation quantities for tablespace and index data sets. Specify too-small amounts, and you might hit the extent limit before reaching the desired maximum data set size (the extent limit is 123 for one volume, and up to 7,257 across 59 volumes for Storage Management Subsystem (SMS)-managed data sets if the z/OS level is at least 1.7; otherwise, up to 255 extents across 59 volumes). Choose too-large values, and you might end up with a good bit of wasted (allocated and unused) space in

Disaster Recovery for AIX



Learn More. Download the White Paper:
visionsolutions.com/IDM1

Data Sharing for DB2



Learn More. Download the White Paper:
visionsolutions.com/IDM2

Call **800-957-4511** or visit **visionsolutions.com** for more information.



Easy. Affordable. Innovative. *Leaders Have Vision.*



your tablespaces and indexes. Multiply this appropriate-allocation-determination exercise by thousands of objects in a huge database, and you've got a big chunk of work on your hands—that is, unless you let DB2 take care of this task for you.

How do you set up DB2 to do that? Easy: start by setting three **ZPARM** parameters as follows:

- ▶ **TSQTY:** 0 (default value)
- ▶ **IXQTY:** 0 (default value)
- ▶ **MGEXTSZ:** YES (default value for DB2 9; for DB2 8 the default value is NO)

Then, for an existing tablespace or index, execute an **ALTER** statement with a specification of -1 for **PRIQTY** and **SECQTY**. For a new object, simply leave the **PRIQTY** and **SECQTY** clauses out of the **CREATE TABLESPACE** or **CREATE INDEX** statement. For the new object (or via **REORG** or **RECOVER** or **LOAD REPLACE** for an existing object),

the primary allocation will be 1 cylinder, and the initial secondary space allocation will also be 1 cylinder. Subsequent extents—up to the 127th—for the data set will be increasingly larger, with sizes determined by a sliding-scale algorithm. Extents beyond the 127th will be of a fixed size based on the maximum size of the data set (127 cylinders for data set sizes up to 16 GB, and 559 cylinders for 32 GB or 64 GB data sets).

Pretty simple, eh? And, it works. When DB2 handles allocation sizing, it's highly unlikely that you'll reach the data set extent limit before reaching the maximum data set size, and the start-small-and-slowly-increase approach to secondary allocation requests keeps wasted space to a minimum.

The bigger the database system...

...the more important it is to use a DBMS that lets you take on "humongous" with confidence. That's DB2. From shared-data

multi-node clustering for super-scalability and ultra-high availability, to advanced autonomies that enable DB2 to manage itself (and the more there is to manage, the more you'll appreciate that), to industry-leading table and index compression capabilities, to a best-of-breed query optimizer that maximizes the bang for your CPU bucks—whether on the mainframe or the LUW platform, DB2 has the technology that enables organizations to effectively manage enormous databases without incurring enormous costs. Put that technology to work for your company, and change information overload into opportunity. *

RESOURCES

IBM DB2 pureScale: ibm.com/db2/9/editions-features-purescale.html

IBM DB2 for z/OS: ibm.com/db2/zos

The best way to perform.

Full DB2 Version 9 support

Integration of DB2PD

24/7 and baseline monitoring

Historic performance data stored in DB2

Screen-driven methodology to facilitate problem-solving

SQL tuning with detection of unused indexes

Gateway Monitor with JDBC Type 4 support

S P E E D  **G A I N**
for DB2

Distributors

USA

ITGAIN Inc.
Phone 1-800 618 1686
1-978-774-8376
Email speedgain@it-gain.com

UK

BLLENHEIM Software Ltd.
Phone +44 (870) 2406771
Email info@blenheim-sw.co.uk

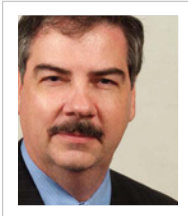
Germany and all other countries

ITGAIN IS GmbH
Phone +49 511 9666 817
Email speedgain@itgain.de

ITGAIN 

Changes to the Cursor Stability Isolation Level: Part 1

Improvements to a familiar feature



Roger E. Sanders

(roger_e_sanders@yahoo.com)
is a consultant corporate systems engineer at EMC Corporation. He is the author of 18 books on DB2 for Linux, UNIX, and Windows and a recipient of the 2008 IBM Data Champion Inaugural award. He is currently working on a new book that outlines how to write technical magazine articles and books and get them published.

Special thanks to Senior Technical Staff Member—DB2 Kernel Architect Mike Winer for providing information used to develop this article.

Isolation levels play a key role in preventing databases from becoming inconsistent in multi-user environments, and of the isolation levels available, Cursor Stability (CS) is probably the one used most often. In this column, I'll show how the CS isolation level worked prior to IBM DB2 9.7, and I'll explain how it differs from the other isolation levels available. I'll also show the lock avoidance techniques that were implemented for the CS isolation level in DB2 9.5.

What is data consistency?

Suppose your company owns a chain of hardware stores and uses a database to keep track of inventory at each store. The database contains an inventory table for each store in the chain, and the table is updated whenever a particular store receives or sells supplies.

Now, suppose a case of hammers is physically moved from one store to another. To reflect this move, the hammer count value in the receiving store's table needs to be raised and the hammer count value in the donating store's table needs to be lowered. But if a user raises the hammer count value in the receiving store's inventory table and fails to lower the hammer count value in the donating store's inventory table, the data will become inconsistent—the total hammer inventory for the entire chain is now inaccurate.

In single-user environments, a database can become inconsistent if a user forgets to make all necessary changes, if the system crashes while a user is in the middle of making changes, or if a database application stops prematurely. In multi-user environments, inconsistency can also occur when several users access the same data simultaneously.

Transactions, isolation levels, and locks

The primary mechanism DB2 uses to keep data consistent is the transaction. A transaction (also referred to as a unit of work) is a sequence of one or more SQL operations grouped together as a single unit. The initiation and termination of a transaction defines points of consistency within a database; either the effects of all operations performed within a transaction are applied to the database and made permanent (committed), or they are backed out (rolled back) and the database is returned to the state it was in before the transaction started.

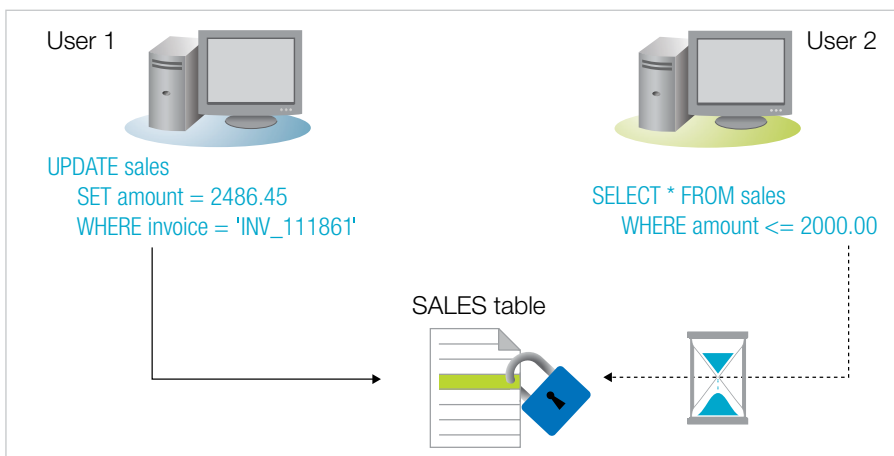


Figure 1: When the Cursor Stability isolation level is used, the row currently referenced by an open cursor is locked—if the row is changed (in this case, by User 1), the lock is held until the change is committed. All other transactions are prohibited from accessing the row until the lock is released (User 2 is forced to wait).

In multi-user environments, transactions often run concurrently. As a result, each transaction has the potential to interfere with other active transactions; the actual amount of interference allowed is controlled by the isolation level used. With DB2, four isolation levels are available:

- ▶ **Repeatable Read (RR):** Prevents dirty reads, non-repeatable reads, and phantoms (see sidebar, “Phenomena seen when transactions are run concurrently”)
- ▶ **Read Stability (RS):** Prevents dirty reads and non-repeatable reads; allows phantoms
- ▶ **Cursor Stability (CS):** Prevents dirty reads; allows non-repeatable reads and phantoms
- ▶ **Uncommitted Read (UR):** Allows dirty reads, non-repeatable reads, and phantoms

Isolation levels are enforced by other mechanisms called locks. Locks are used to associate a data resource with a single transaction, for the sole purpose of controlling how other transactions interact with that resource while it's associated with the transaction that locked it.

The Cursor Stability isolation level

The CS isolation level is used much more frequently than the other isolation levels because it provides the greatest amount of concurrency while preventing dirty reads. SQL statements running under this isolation level acquire a lock only for the row that is being referenced by an open cursor. Once acquired, this lock remains in effect until the cursor is repositioned or closed or until the owning transaction is terminated.

If the cursor is repositioned, the lock held on the previous row is released (provided no data changes were made), and a new lock is acquired for the row where the cursor is now positioned. (The cursor position applies to the internal cursor on the server—not necessarily the cursor seen by an application.) If data in the original locked row changed, the lock is held until the change is committed (see Figure 1).

Transactions using the CS isolation level do not see uncommitted changes made by other transactions. However, they can see non-repeatable reads and phantoms.

Lock avoidance techniques

Prior to DB2 9.5, if the CS isolation level was used and a row was locked on behalf of a transaction, DB2 would block attempts by other concurrently running transactions to modify the locked row. If the locked row was changed in any way by the transaction holding the lock, other SQL statements in concurrent transactions were not allowed to access the locked row (unless they were running under the UR isolation level) until the transaction was committed: writers would block readers and in some cases readers could block writers. In either case, concurrent transactions that needed access to a row that was locked were forced to wait for the lock to be released before they could continue processing. This behavior could cause lock waits, lock timeouts, or deadlocks to occur.

To eliminate some of the locking overhead required for the CS isolation level, a number of lock avoidance techniques were introduced in DB2 9.5. These techniques allow scan operations to execute without locking rows when the data and/or pages being accessed are known to be committed. For example, consider the following query:

```
SELECT COUNT(*) FROM sales
```

Prior to DB2 9.5, if this query was executed, the first row in the SALES table would be locked, a count would be taken, and the lock would be released. Then the second row in the table would be locked, the count would be updated, and the lock would be released. This behavior would continue until all of the rows in the SALES table had been counted.

Starting with DB2 9.5, the same query would result in the rows being counted, but the locks would no longer be acquired and released—provided DB2 can determine that the rows are committed without acquiring locks. Essentially, lock avoidance allows DB2 to determine whether the data needed is committed—if it is, locks are not acquired. With DB2 9.7, lock avoidance works for any read-only SQL statement running under the CS isolation level that is using cursor blocking. Cursor blocking is a technique that reduces overhead by having the database manager retrieve a block of rows in a single operation.

Currently Committed semantics

DB2 9.7 introduces a new implementation of the CS isolation level that incorporates Currently Committed (CC) semantics to further prevent writers from blocking readers. In my next column, I'll introduce you to CC semantics and I'll show you how they provide faster data access and increased data concurrency for SQL statements running under the CS isolation level. *

PHENOMENA SEEN WHEN TRANSACTIONS ARE RUN CONCURRENTLY

When transactions are run concurrently, four types of phenomena can occur:

- ▶ **Lost updates:** Two transactions read the same data and both attempt to update that data, resulting in the loss of one of the updates.
- ▶ **Dirty reads:** A transaction reads data that has not yet been committed.
- ▶ **Non-repeatable reads:** A transaction reads the same row of data twice and gets different results, usually because the data was changed by another transaction.
- ▶ **Phantoms:** Phantoms are results that aren't returned by the first execution of a search query, but they appear in subsequent executions. Phantoms usually consist of data that has been added or changed between the time that the queries are run.

Fastest Informix DBA Contest II: How Did They Do It?

The winners of the latest contest tuned a process from 40 hours to 1 minute. Here are their techniques.



Lester Knutsen

(lester@advanceddatatools.com) is president of Advanced DataTools Corporation, an IBM Informix consulting and training partner specializing in data warehouse development, database design, performance tuning, and Informix training and support. He is president of the Washington, D.C. Area Informix User Group, a founding member of IIUG, an IBM Gold Consultant, and an IBM Data Champion.

Over the summer, we ran another Fastest Informix DBA contest based on our very successful contest at the IIUG Informix Conference in April 2009. We enhanced the rules, made the benchmark process harder, and doubled the size of the database: a customer billing process that had lots of unnecessary SQL, missing indexes, and an `ONCONFIG` file with some really bad configuration settings. The billing process took 40 hours to complete, and we challenged the participants to make it run faster.

We had more than 70 participants, of whom 8 tuned the process to run in less than 6 minutes. The fastest made it run in a little under 1 minute. And to add to the excitement, only a 5-second difference separated the top three places.

So, congratulations to the new winners! The results were announced at the IBM Information On Demand 2009 Global Conference and updated in a Webcast on November 16. I made a mistake and missed the fastest entry, so we decided to give out two grand prizes. To qualify for the grand prize, the user must be a DBA employed at a company using Informix internally, not a consultant, and not an IBM employee.

- ▶ Grand Prize and Fastest Overall Time—Fastest User DBA: Tatiana Saltykova
- ▶ Grand Prize—Fastest User DBA: Eric Rowell
- ▶ Fastest Consultant: Warren Donovan
- ▶ Fastest International DBA: Malte Sukopp, Germany
- ▶ Runner-up Consultant: Jeff Filippi

- ▶ First Runner-up User DBA: Yunyao (Frank) Qu
- ▶ Second Runner-up User DBA: Tammy Frankforter
- ▶ Fastest Youngest DBA: Riya Kariath

In this column, I want to highlight what Tatiana, Eric, and Warren did to take a 40-hour process and make it run in 1 minute. They all highlight great examples of what a DBA must do to produce fast code and fine-tuned databases.

Create the right indexes

The database had four tables, each with a primary key, but no other indexes. One of the tricks to database performance is identifying the right number and placement of indexes. Missing indexes will slow down reads, but too many indexes will slow down inserts, updates on indexed fields, and deletes.

The billing process was missing one key index. After the bills were created, the customer table was updated with a new balance, which required the bills table and customer table to be joined (see Figure 1). However, while the customer table had an index on the customer number field because it was the primary key, the bills table did not. Without this index, the only way to find a customer was to do a sequential scan of more than 600,000 bills. Updating 101,000 customers would require 101,000 x 605,280 bills or 61,133,280,000 scans of the bills table. Simply adding this index on the customer number field in the bills table reduced the processing time from 40 hours to about 30 minutes.


```

update customer
  set balance_due = balance_due + ( select sum ( total_bill )
    from bills where bills.customer_number = customer.customer_number )
  where customer_number in ( select customer_number from bills );

```

Figure 1: SQL to update a customer balance—requires an index on customer_number

```

CREATE TRIGGER ins_bills insert on bills
  REFERENCING NEW AS n
  FOR EACH ROW
  (UPDATE customer
    SET balance_due = balance_due + n.total_bill
    WHERE customer_number = n.customer_number);

BEGIN WORK;
LOCK TABLE bills IN EXCLUSIVE MODE;

-----
Create bills
-----

insert into bills
  (
    customer_number,
    last_name,
    ...
    total_bill
  )
select
  customer.customer_number,
  customer.last_name,
  ...
  product.product_price,
  CASE
    WHEN customer.start_date <= "01/01/2009"
      AND customer.balance_due > 50000
    THEN 10
    ELSE 0
  END,
  state.sales_tax,
  CASE
    WHEN customer.start_date <= "01/01/2009"
      AND customer.balance_due > 50000
    THEN ((product_price - 10) * (1 + state.sales_tax))
    ELSE (product_price * (1 + state.sales_tax))
  END
from customer, state, product
where customer.state = state.state
and customer.product_code[1] = product.product_code[1]
and product.product_number in (1, 2, 4, 7, 9, 10);

```

Figure 2: Eric Rowell's Fastest Informix DBA SQL. This SQL takes three inserts and two update statements and optimizes them into one SQL statement and one trigger, increasing performance.

It's also important to add only necessary indexes. Several participants created additional indexes that did not help the Informix SQL optimizer and actually slowed down the process.

Optimize the SQL statements

The benchmark process had five SQL statements: three inserts into the bills table and two update statements—one to calculate the total bill discount, and one to calculate the new balance. Eric optimized the SQL to a single statement by using a trigger that updated the customer table only when a bill was inserted into the bills table (see Figure 2). This is very efficient code and one of the reasons he performed the task so fast. Eric eliminated the bill discount `UPDATE` by adding two SQL `CASE` statements to the `INSERT` statement. Meanwhile, the trigger on the bills table made the new balance `UPDATE` statement unnecessary because the customer balance was updated anytime a bill was inserted.

Eric's trigger was an especially brilliant approach because it also eliminated the need for the index on the customer number field in the bills table previously discussed. He eliminated the time to create and maintain this index because he did not need it, and he got the job done faster.

The key to optimizing your SQL is to reduce the number of statements that read through tables. The baseline SQL in this contest read the customer table three times, one for each of the `INSERT` statements. And, it read the customer table and the bills table two more times for each of the updates. By reducing the SQL to one read and a trigger that immediately updated the customer table, Eric cut the disk I/O and database reads and writes to one-fifth of the baseline.

Reduce the disk I/O

Another technique that both Warren and Eric used was to reduce the number of disk reads. Both used a new feature of Informix: creating `dbspaces` with a 16 KB page size instead of the default 2 KB page size. This meant that the database engine could gather eight times more data in one read, which benefits most indexes by putting more index data on a single page. The benefit lies in reducing the number of time-consuming disk reads to get all the data into memory. Warren and Eric both set up most of their buffer pools for the 16 KB pages, so this optimized the amount of work that could be done in memory.

More to come at the 2010 IIUG Informix Conference

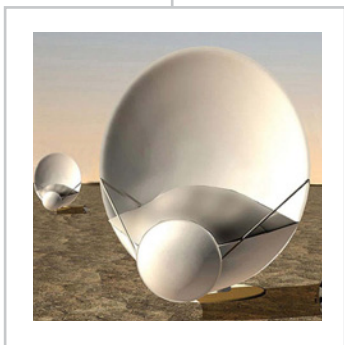
The contest was a lot of fun to run and monitor, and it is exciting to see the ingenuity and creativity that all the Informix DBAs put into it. Congratulations to all the winners, who are listed on our Web site: www.advanceddatatools.com/Informix/index.html.

We will sponsor the next version of this contest, the Fastest Informix DBA Contest III, at the IIUG Informix Conference in Overland Park, Kansas, April 25–28. Visit the Advanced DataTools Web site above for more details or the International Informix User Group Web site at www.iiug.org/conf. Hope to see you there. ✧

Smarter is...

Understanding the Universe

IBM is teaming up with scientists in Australia to tame exabytes of data from space



Chris Young is a technology writer based in the Pacific Northwest.

Building one of the largest scientific instruments on the planet comes with an equally large challenge: how to handle the astronomical amounts of data that the new instrument will generate.

IBM and Western Australia's International Centre for Radio Astronomy Research (ICRAR) plan to answer that question. Scientists at ICRAR are working to create the Square Kilometre Array (SKA), a next-generation radio telescope that will use a huge array of antennas to give astronomers insights into the evolution of galaxies, dark matter, and energy.

Far more sensitive than today's instruments, SKA will produce vast quantities of data when it becomes operational between 2020 and 2025: one exabyte—a thousand million gigabytes—every 24 hours. This amount is too large and too continuous to place directly into memory in a cost-effective manner. Instead, the data will be processed by a pair of supercomputers designed to refine the raw data and convert it into images. According to ICRAR Director Professor Peter Quinn, "This task will involve developing concepts for technology that, before the SKA, had never been needed."

ICRAR and IBM are already exploring new ways to transfer, process, and store the unprecedented

amounts of data. IBM experts are currently studying types of solid-state storage, high-speed fiberoptic transmission, and machines with the processing power of approximately one billion PCs.

To answer questions about the universe, scientists will also need database capabilities unlike anything available today. "The SKA images will be stored as three-dimensional data objects, or cubes, with length, breadth, and radio frequency information," says Quinn. "Each cube can be tens or hundreds of terabytes in size, and a useful image of the sky could contain hundreds of these cubes. So we will need to develop powerful new data discovery engines and databases that can hold a billion objects."

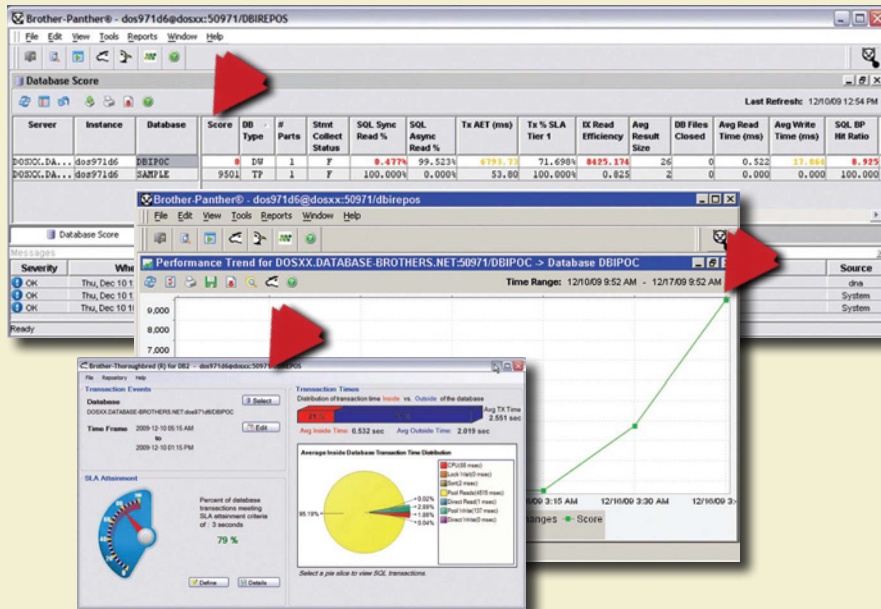
Quinn is confident the ICRAR and IBM partnership will push the boundaries of data management and contribute to making a smarter planet. "IBM is a world leader in the field of technology development, and its expertise will help us develop tools to handle the colossal amounts of data at the speed the SKA demands," says Quinn. "That work will apply to a host of other areas where real-time management of vast, complex data promises to be key, from holographic imaging in healthcare to large sensor networks for examining the ocean depths." *

Manage Datacenter Productivity and Profitability

Your business relies on data management to be successful. Whether you use datacenters to handle transactions or as repositories for information critical to decision-making, speed of response and continuous availability is essential to your success.



Do you know what this speed and availability costs?



DBI tools for DB2 LUW give you instant access to database performance analysis, suggest best practice improvements and measure results. The top screen shows a database score – like a credit score – and meaningful KPIs to guide you to improved performance.

DBI'S INNOVATIVE SOLUTION

Standard solutions can chip away at profitability. **Adopt DBI's innovative, methodology.** DBI tools benchmark database and datacenter performance, quantify, aggregate total operational costs measured against customer experience and service level attainment, and implement datacenter best practices that ensure **continuous improvement** and **customer satisfaction**.

DBI Tools Quickly Improve Datacenter Profitability

- Reduce total operational cost
- Increase productivity
- Improve customer service
- Reallocate resources
- Repurpose existing personnel and facilities

DATACENTER BEST PRACTICES

Get facts – don't make decisions based on anecdotal or intuitive sources. DBI tools quickly provide objective information:

- Measure datacenter performance
- Isolate performance bottlenecks
- Enact solutions
- Measure again to quantify improvement



Datacenter performance is not immutable – continuous evaluation is needed to maintain peak performance and productivity.

Measure, improve, verify, repeat.

Reduce Total Operational Costs...

Your bottom-line success is impacted by these aggregated costs:

- Tuned and optimized databases consume **less server capacity**
- Datacenters which aren't optimized face a perpetual need to **add expensive hardware** to handle demand
- Expanded hardware in datacenters means you'll be required to purchase **additional software licenses**
- New hardware requires **larger facilities**
- Larger facilities drive up **utility costs**
- **Additional personnel** are required to manage expanded datacenters
- Tuned and optimized databases consume **22-44% less electricity**
- Average datacenters spend **55% of their energy budget on cooling**

Are you migrating to DB2 LUW from Oracle? Get an expert, independent, objective opinion on your performance readiness.



Contact DBI Today For More Information

Phone: (512) 249-2324

Toll-free: (866) 773-8789

www.DBISoftware.com/profitability



TDA GROUP
800 El Camino Real, Suite 380
Mountain View, CA 94040
U.S.A.

Sponsors of Tomorrow.™ 

IT sees:

Servers that intelligently
scale performance.

The CFO sees:

Servers that save energy.



The latest Intel® Xeon® processor analyzes its workload and automatically adjusts to deliver maximum performance when you need it—and big energy savings when you don't. That's smart no matter how you look at it.

Learn more at intel.com/go/xeon.