Redbooks

ibm.com/redbooks

SIMD Business Analytics Acceleration on z Systems

Rajaram Krishnamurthy

IBM Academy of Technology

Learn

IBM z Systems

IBM

Point-of-View

SIMD Business Analytics Acceleration on z Systems

An IBM Redbooks Point-of-View Publication from IBM Systems

By **Raj Krishnamurthy**, Sr. Tech.Staff Member

Highlights

- The new Single Instruction Multiple Data (SIMD) accelerator is integrated into the latest IBM z Systems™ CPUs to enable workloads for acceleration by using new types of instructions that can act on multiple data elements simultaneously, even at the level of a single instruction.
- SIMD acceleration can help IT organizations meet shrinking batch window times and increasing data throughput requirements for mainframe workloads. It can also be applied to extract real-time insight from financial and consumer transactions because with acceleration, the workloads can keep up with incoming transaction rates.
- The SIMD accelerator and software stack allow workloads from previous generation systems to run unchanged on the new z Systems products enabled for SIMD acceleration. Workloads can also be modified and enabled for higher gains with SIMD acceleration.

Analytics workload landscape

For organizations that use them, IBM® z Systems mainframes are the principal source of operational data in the enterprise. And obtaining real-time insight from this operational data provides a competitive advantage. Examples include real-time fraud analytics, near real-time trade settlements, determining next-best actions for customer service, and so on.

The ability to turn data into accurate and timely insight is determined by the efficiency of your business analytics execution and the relative proximity of the operational data to be analyzed. In addition, analytics workloads must be able to process a variety of data types, including numerical integer, numerical floating-point, and character string data.

The SIMD accelerator is integrated into the newest z Systems CPUs so that analytics workloads can act on multiple data items simultaneously, using a single instruction. This can enable z Systems-based analytics, which by their nature run in near proximity to the operational data, to keep up with incoming transaction rates, even when asked to deliver real-time insight.

Figure 1 shows the combined impact on workloads of compute intensity (compute throughput measured in operations per second) and data intensity (data throughput measured in bytes per second).

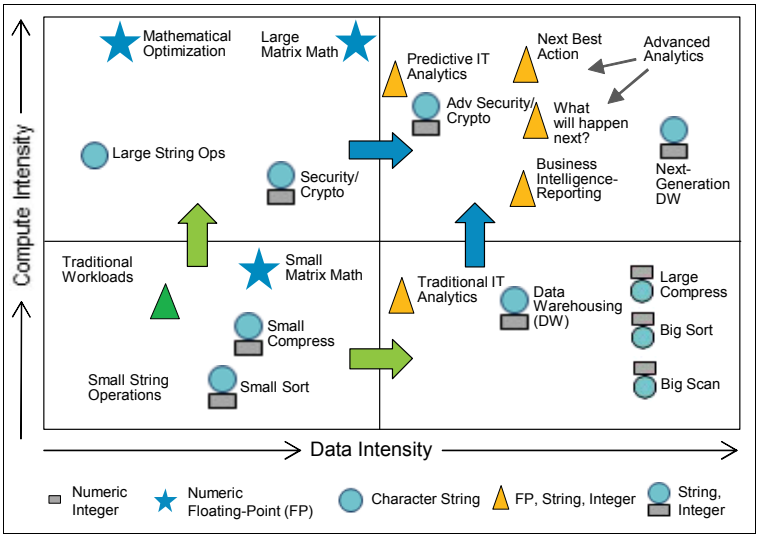

Figure 1 SIMD can benefit data- and compute-intensive workloads

As available time windows for batch processes shrink to meet shifting global business requirements, workload compute throughput must increase to fit within the smaller windows. Data processing throughput must increase as well to handle ever-higher data volumes and data velocities.

In Figure 1, traditional workload processing is performed in the lower left quadrant for character string-based workloads (such as string sorts) or numeric workloads (such as small matrix mathematical calculations). With shorter batch time windows available, workload compute throughput requirements move (increase) into the upper left quadrant (indicated by the vertical green arrow). The SIMD accelerator can be engaged to process workloads to meet the higher compute throughput requirements in this quadrant. Workloads that can benefit from SIMD acceleration in this quadrant include string processing-intensive workloads, security and cryptographic workloads, and mathematical modeling workloads (such as complex business optimization problem solving).

As data volumes and velocities increase, workloads in the lower left quadrant of Figure 1 on page 1 become subject to the data throughput requirements of the lower right quadrant (indicated by the horizontal green arrow). The SIMD accelerator in the new IBM z13™ has special features to process various character string, numeric floating-point, and numeric integer data at the necessary throughput rates. Workloads that can benefit from SIMD acceleration in this quadrant include next-generation data warehousing and IT analytics.

Advanced analytics workloads enabled by the z13 platform are shown in the upper right quadrant of Figure 1. Examples include prescriptive analytics, such as for determining the next-best action in a customer service scenario, and predictive analytics, such as fraud detection and advanced data warehousing. These workloads process character string and numeric data, which makes them ideal for SIMD acceleration.

Over time, the compute and data processing throughput requirements for workloads in the lower left, upper left, and lower right quadrants are likely to increase. They might overlap with those for workloads in the upper right quadrant.

When a business uses z Systems mainframes as their business analytics hub and consolidates business analytics processing there from various lines of business, the SIMD accelerator can provide benefits across the enterprise and for a variety of data sources, both structured and unstructured, with the data at rest or in motion.

The SIMD software stack: Enabling workloads for SIMD

Figure 2 shows the SIMD software stack.

In what we call *Transparent* execution mode, workloads from previous generation systems can be run directly on the z13 to gain SIMD acceleration benefits without explicit programming of the SIMD accelerator. For example, Java workloads with string and character processing content can be seamlessly migrated to the z13 using Java8 (the next major version of Java) and can be enabled to benefit from SIMD acceleration. Also in *Transparent* execution mode, workloads using IBM or ISV software products, when migrated from previous generation systems and run unmodified on z13, can also be enabled to benefit from SIMD acceleration.

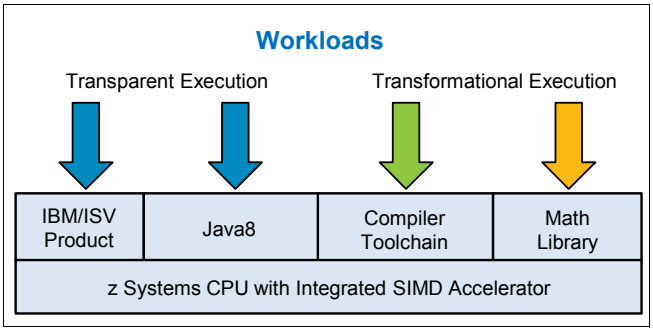

Figure 2 The SIMD Software Stack

In so called *Transformational* execution mode, workloads are enabled to benefit from SIMD acceleration by the IBM Compiler and Language Environment®, through either source code transformations, executable rebuilds, or both. Depending on the language environment (C/C++, COBOL, or Assembler), SIMD-optimized software libraries might need to be used to enable SIMD acceleration of workloads in *Transformational* execution mode.

Benefits of SIMD acceleration

The SIMD Accelerator can provide several key advantages. Here is a partial list:

- Run analytics in close proximity to z Systems data, with high performance and efficiency, to meet new, more demanding requirements for batch processing windows, data processing rates, and data volumes and velocities
- Enhance the accuracy and timeliness of real-time business insight

- Enable faster mathematical modeling solutions to complex business optimization problems to meet both real-time and batch time processing windows
- Accelerate Java workloads that have string-rich and character-rich content
- Allow construction of richer, more complex analytics modeling workloads to provide better accuracy of insight
- Allow analytics workloads to be easily moved from existing platforms to z Systems to provide faster business insight
- Increase programmer productivity when developing analytics workloads, so business insight can be generated even faster and greater competitive advantage can be achieved

SIMD Use Case 1: Real-time fraud detection

Figure 3 shows the steps used in detecting fraud in the transfer of funds between financial accounts.

The first step is building an analytical model, typically based on several months of consumer transactions. Within model building, the initial step is data curation, in which an analyst explores the data to determine what might be learned from it. This exploration is highly data parallel in that multiple data elements can be operated on simultaneously, even at the level of a single workload instruction. After a useful set of data is chosen, this data can be used to *train* the model.

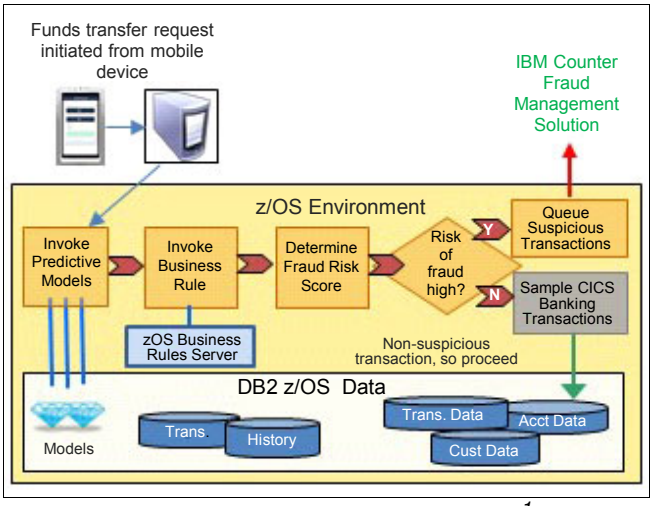

Figure 3 Steps used in real-time fraud detection[1]

[1] Adapted from *System z Integrating Transaction and Analytics for Fraud Detection – Extended* (video), Mythili Venkatakrishnan and Michael Poirier (IBM). See Resources section for URL.

Data clustering, a type of machine learning algorithm, is a typical step in model training where data points sharing common characteristics are grouped in clusters. This step is also highly data parallel because multiple clusters can usually be identified from the data points independent of each other. In general, various machine-learning algorithms are easily accelerated using the SIMD feature because of their data parallel compute behavior.

After the model is built (see the diamonds in Figure 3), it can be published into a user-defined function (UDF) or placed in other software objects for later execution. This allows models to be run to recognize risk based on data from transactions. As transactions arrive into the system, the process of recognizing risk in a transaction in real time is also amenable to SIMD acceleration.

When a transaction enters the system (when a funds transfer request is submitted), the process of extracting information from fields and analyzing the data is highly amenable to SIMD acceleration. Examples of these fields include the physical location of the person requesting the transaction, the amount of the transaction, and the source and destination accounts. After the required data is extracted, the predictive model is run and a score can be determined.

The process of scoring is also highly suitable for SIMD acceleration and can be done in parallel fashion. During scoring, the difference between the features of each incoming transaction and what the model expects them to be can be represented numerically and computed. After scoring is completed, the extracted transaction data is passed through business rules engines, which can also benefit from SIMD acceleration. *If-then-else* rules can be transformed into lookup tables with bit patterns, after which SIMD accelerator instructions for bit manipulation can be used for business rule execution. A composite fraud risk score can be computed by using the combined operation of the model and execution of business rules. This process is also ideal for SIMD acceleration because the computation usually involves vectors of numeric fields. If the risk score is lower or higher than a predetermined threshold, the transaction can be submitted for further processing. Otherwise, it can be flagged as a suspicious transaction and be sent to a different system for more detailed fraud analysis.

SIMD acceleration for fraud detection can speed the production of accurate models by running model training with higher throughput. SIMD acceleration can also improve the timeliness of real-time fraud

determination because it enables models and business rules to be evaluated with higher efficiency.

Use Case 2: Next-best action for customer service

A so-called *next-best action* advanced analytics system captures customer activity from various input channels such as mobile apps, email, text, chat, and telephone interactions. It then recommends solutions to resolve a customer's specific problem or improve the customer's overall relationship with the organization. "Next-best" refers to an analytics system's ability to provide a customer service agent with a series of automated potential solutions to a customer problem.

Figure 4 shows a typical next-best action system.

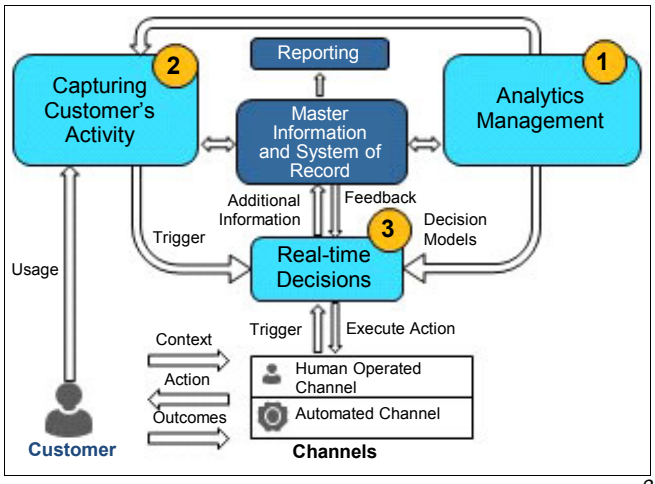

Figure 4 Determining the next-best customer service action[2]

In the Analytics Management block, labeled 1, decision models are created by data analysts who curate the customer activity data to identify patterns. This discovery step is highly parallel, and therefore can benefit from SIMD operations on the involved integer, string, and floating-point data. The curated data is then used to train decision models, a process that, as explained earlier, is also highly amenable to SIMD acceleration. This step usually builds models that can be incrementally updated or models that continuously learn from execution in the field. At this stage, business rules are often applied.

The execution of business rules is also suitable for SIMD acceleration in that it often relies on using bit-pattern lookup tables created directly from if-then-else rules. After the decision model is prepared, it can then be injected into appropriate software assets for deployment. Model preparation usually involves processing several structured and unstructured data sources and, as such, involves numeric integer, numeric floating-point, and character string data that are appropriate for SIMD acceleration.

Customer activity data is usually a combination of structured and unstructured data (labeled 2 in Figure 4). The process of extracting data from incoming sources is also highly parallel, and SIMD acceleration can be applied to several of the required data permutation and shuffling operations. As customer activities are logged, overall progress towards business objectives (such as increased ROI, and increased upsells and cross-sells) can be recorded and can trigger decision models (and related actions) to be run. Several functions in these steps are suitable for SIMD acceleration because they involve processing multiple data elements simultaneously. When a customer interacts with the business, the context of the interaction (for example, the precipitating event, product, or service they are calling about) can be transmitted across an input channel. If the interaction of the customer with the system generates a trigger, such as when a customer expresses dissatisfaction with current services, then further action might be needed. So the trigger is communicated to the Real-Time Decisions engine (labeled 3 in Figure 4) where the decision model is run with data inputs extracted from the interaction context. Upon execution of the model, the engine communicates a recommended action to the customer service agent for any remedial measures.

Several functions in this process are inherently data parallel, as they involve processing independent fields. The model execution is also amenable to SIMD acceleration because it involves data parallel machine-learning algorithms such as classification and distance vector computation. The need for processing several independent contexts at the same time, coupled with the diversity of data (numeric integer, numeric floating-point, and string) make next-best action systems ideal for SIMD acceleration.

SIMD acceleration for next-best action systems can improve an organization's ability to produce more accurate decision models in ever-shrinking time windows with higher throughput. It can also provide the ability to make real-time decisions faster, even *during* customer interactions.

[2] Adapted from Smarter Analytics: Driving Customer Interactions with IBM Next-Best Action Solutions (IBM Redbooks® publication), Mandy Chessell and David Pugh. See Resources section for URL.

What's next: How IBM can help?

The SIMD Accelerator is available exclusively on z Systems, starting with the new z13 platform. It can help accelerate a variety of workloads, even beyond those listed in this Point of View, and particularly those that are business analytics-oriented or rich in compute- or data-processing requirements.

IBM will engage with ISVs to enable their products for SIMD acceleration. IBM products that are enabled for SIMD acceleration will be highlighted when those products are released. And IBM will create documentation about SIMD, including a new publication, *The SIMD Accelerator for Business Analytics on the z13*[3], to be published later in 2015. And new SIMD instructions will be documented in the next release of *z/Architecture Principles of Operations* (ideal for assembly language programmers or to gain a deeper understanding of SIMD).

Customers can upgrade to the appropriate version of IBM products on z13 to enable their workloads for SIMD acceleration. To experience the benefits of Transformational execution with C/C++ workloads, as shown in Figure 3 on page 3, customers can upgrade to the appropriate XL compiler release on z13 to use SIMD built-in functions and math libraries.

In addition, IBM client teams can be used to find SIMD programming resources for assistance with source code optimizations for SIMD acceleration. And IBM plans to enrich other language environments, such as PL/1 and Enterprise COBOL, with SIMD acceleration support.

Resources for more information

- IBM Mainframe Business Analytics and Data Warehousing:

  http://www-01.ibm.com/software/os/systemz/badw/

- IBM *System z* - Real-time analytics for operational data:

  http://www-03.ibm.com/systems/z/solutions/data.html

- Big data on System z:

  http://www-01.ibm.com/software/data/bigdata/z/

- IBM XL C and C++ Compilers family:

  http://www-03.ibm.com/software/products/en/ccompfami

- IBM Mathematical Acceleration Subsystem family:

  http://www-03.ibm.com/software/products/en/mathaccesubsfami

- Java Standard Edition Products on z/OS:

  http://www-03.ibm.com/systems/z/os/zos/tools/java/

- *System z Integrating Transaction and Analytics for Fraud Detection – Extended* (video), Mythili Venkatakrishnan (IBM) and Michael Poirier (IBM):

  http://youtu.be/CWQNJ2ystic

- *Smarter Analytics: Driving Customer Interactions with IBM Next-Best Action Solutions* (IBM Redbooks publication), Mandy Chessell and David Pugh

  http://www.redbooks.ibm.com/abstracts/redp4888.html?Open

[3] E. M. Schwarz, R. B. Krishnamurthy, C. J. Parris, J. D. Bradbury, I. M. Nnebe, and M. Gschwind, "The SIMD Accelerator for Business Analytics on the z13," IBM J. Res. & Dev., vol 59, no. 4/5, 2015 (to appear later in 2015).

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

This document, REDP-5145-00, was created or updated on February 5, 2015.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (or), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at `http://www.ibm.com/legal/copytrade.shtml`

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:
IBM®
Language Environment®
Redbooks®
Redbooks (logo) 
System z®
z/Architecture®

The following terms are trademarks of other companies:

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Other company, product, or service names may be trademarks or service marks of others.