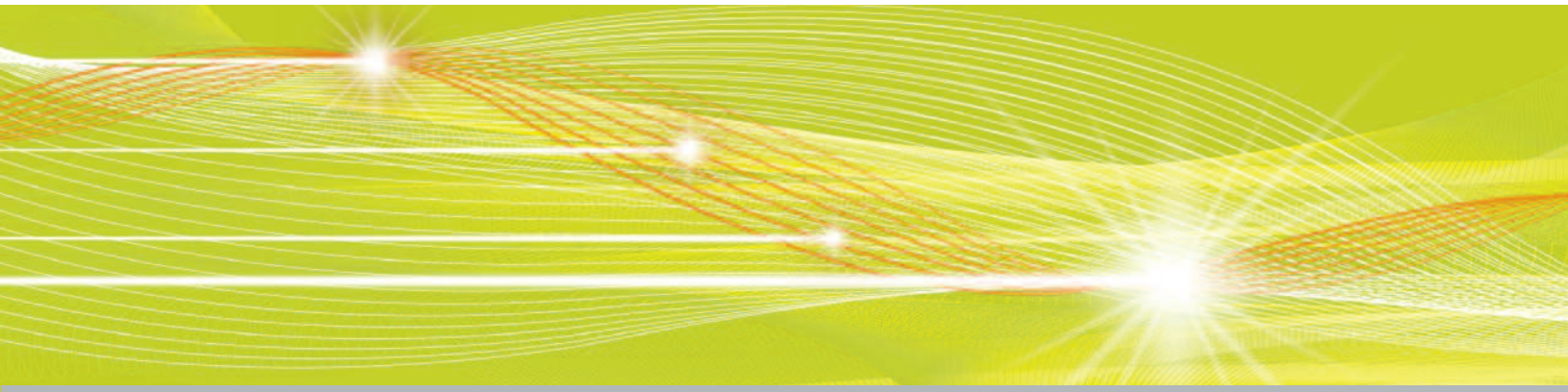
NETEZZA

The Power to Question Everything™

Streaming Analytics™ with the Netezza Performance Server® Appliance

Bringing Appliance Simplicity to Advanced Analytics

Whitepaper



Introduction

As the pioneer in appliances for analytic processing, Netezza recognized years ago that operating at streaming speeds can be game-changing for data-intensive companies. The architecture of its Netezza Performance Server® (NPS®) system is based on this principle, improving performance by orders of magnitude. Netezza customers around the world and across many industries are reaping the benefits – wherever critical information is needed urgently.

Initially developed for SQL queries, the NPS system is now enabled for high-performance analytics, allowing algorithms to run 10-100 times faster than traditional approaches. Wall Street firms can run complex risk analyses iteratively, hundreds of times throughout the day rather than overnight, for better trading decisions. Telecommunications companies can analyze call data records (CDRs) in real time to offer customers the best packages and rates. Retailers can optimize SKU-level pricing by capturing dynamic in-store data. The NPS system has a role in many organizations where fast, in-depth analysis of data impacts business performance.

This paper examines how the Netezza family of appliances is uniquely suited for running complex analytic applications on massive data sets. We begin by addressing streaming processing – the "secret sauce" within Netezza's architecture. We'll then see how that groundbreaking architecture works, and how Netezza innovations bring the power of streaming analytics™ to our customers

Streaming Processing and FPGAs

Many applications require high-performance processing of data streams, where massive amounts of image or signal data flowing into a system must be processed immediately. Streaming processing allows you to watch videos on your computer, and is also used in industrial controls, medical imaging, telecommunications devices, missile guidance systems and many other applications.

Specialty chips, sometimes called hardware accelerators, are used to handle streaming processing tasks. The chip performs early-stage processing of the data stream – initial filtering, correction and conditioning – before passing the refined data to an Intel microprocessor, Power PC or some other device for further processing. Using this approach, a balanced architecture can be designed that combines accelerator technology and general-purpose processors to best fit the application requirements.

Different types of hardware accelerators are available, the most common of which are Application-Specific Integrated Circuits (ASICs) and Field Programmable Gate Arrays (FPGAs). While customized ASICs are used in some streaming applications, their inherent inflexibility, long design cycles and associated cost can make them a less than ideal solution. FPGAs provide a better alternative for many high-performance requirements, providing reconfigurable flexibility and commodity pricing while avoiding the long design cycles typical of ASICs.

How FPGAs work

An FPGA is a semiconductor chip with a series of internal gates that can be programmed to implement almost any logical function. Data enters the FPGA and passes through the gates according to the function loaded into the chip. Filtering and processing takes place as data flows unimpeded through the gates. This accounts for the speed advantage of FPGAs when handling data streams – in contrast, conventional x86-based microprocessors from Intel, AMD and others typically require shuttling data between storage and memory to perform a processing operation.

FPGAs can be issued new instructions for different tasks – hence the term "field programmable." With gate counts of over one million, modern FPGAs can implement much of the functionality in today's systems. These large gate counts and the reconfigurable nature of FPGAs make them an attractive component for a growing range of streaming applications.

FPGAs are everywhere

The FPGA has blossomed in recent years to take on a key role in driving cost-effective high-performance in many different applications. In consumer electronics, for example, FPGAs help process audio and video data for DVD/MP3 players, plasma displays, HDTV and a growing number of exciting new products. They are also used extensively in industrial process controls to filter streaming data from sensors and instruments. In medical imaging systems, FPGAs filter incoming data and also perform processing-intensive image reconstruction. They're behind the dashboard of our cars, where they are widely used in performance, navigation and infotainment systems. And as conventional microprocessors show signs of approaching their performance limits, leading computer companies and research labs are experimenting with FPGAs for a wide variety of supercomputing applications.

Netezza's Innovation: Programmable Streaming Processing for Analytics

Netezza's insight was recognizing that data streaming using FPGAs could play a key role in data warehousing and analytic applications, where enormous amounts of data have to be examined and processed. This understanding was key to development of the first NPS system, an innovative data warehouse appliance that optimizes the use of several processing technologies for huge performance gains – 10-100 times faster than traditional, general-purpose systems. More than 100 Netezza customers are now using Netezza appliances to run complex SQL queries against terabytes of data. But beyond SQL, Netezza is broadening the use of its architecture and processing capabilities to address a wider set of analytic challenges, including those outside traditional database processing.

The appliance architecture is based on a fundamental computer science principle: ***when operating on large data sets, do not move data unless you absolutely have to***. The NPS system fully exploits this principle by processing data extremely efficiently, as early in the data stream as possible. By optimizing the use of FPGAs and other commodity components, the architecture has revolutionized the data warehousing industry, enabling Netezza to deliver tremendous performance in a compact, low-power appliance that is fast to install and incredibly simple to operate.

Netezza architectural overview

From an architectural standpoint, the NPS system consists of two tiers: an SMP host and hundreds of massively parallel blades that Netezza calls Snippet Processing Units (SPUs). The host compiles application commands, assigns processing tasks to the SPUs and returns results to the user. The SPUs are responsible for processing individual query segments ("snippets") on their portion of the database. Each SPU consists of a CPU, memory, a disk drive and an FPGA chip that filters records as they stream off the disk. Because the FPGA handles up to 90% of SPU processing, the CPU used in each SPU is a low-cost Power PC – more than adequate for remaining processing tasks.

When a user submits a SQL query, the host compiles it and generates an execution plan that identifies which tasks need to be performed at the host and which should be executed in parallel by the SPUs. The host then divides the parallel processing tasks into a series of snippets and distributes the snippets to the SPUs for execution. Each snippet is expressed as compiled code for an SPU's CPU together with parameters for its associated FPGA for filtering and projecting data. With this approach, the NPS system optimizes the hardware and software for highest performance for each snippet executed.

Using this Netezza architecture, processing tasks are handled with elegant simplicity. The FPGA performs initial filtering and processing functions as quickly as data can stream off the disk, and passes the results to its associated CPU. The CPU performs any additional processing tasks and passes its results to the host. The host assembles individual results from the SPUs, performs any final processing required and returns the final results to the user.

These multiple levels of optimization in hardware and software filter out extraneous information as early in the data flow as possible, greatly reducing the processing burden downstream. This approach eliminates the I/O bottlenecks that occur when general-purpose processing architectures are used to examine massive amounts of data, making the NPS system dramatically faster than conventional systems.

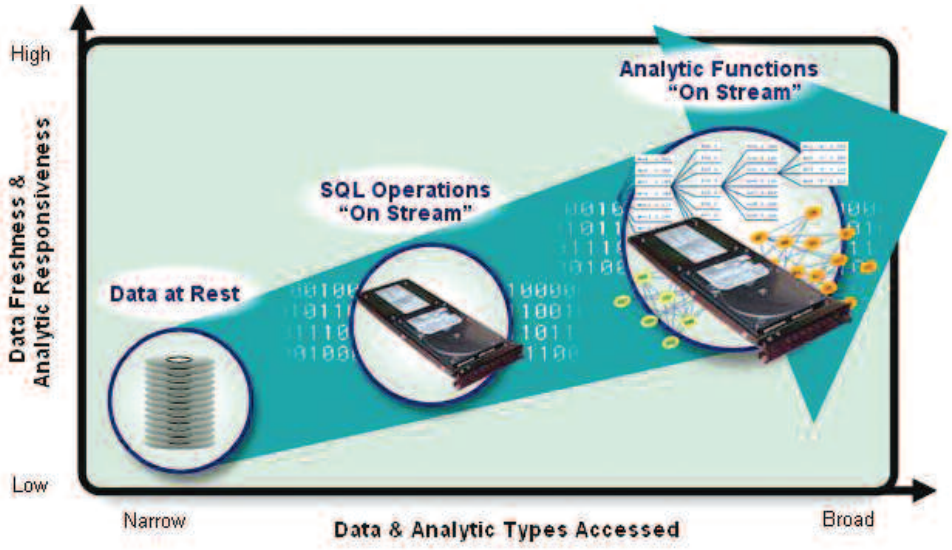
From Streaming SQL to Streaming Analytics

By "bringing queries to the data," Netezza has transformed the world of data warehousing. Now Netezza is extending the capabilities of its NPS system by "bringing analytics to the data": embedding complex non-SQL algorithms used by analytics applications within the Netezza appliance so they can be run "on stream."

Until now, many types of analytic processing were impossible within a database. Complex analytic problems that could not be expressed in SQL had to be solved outside the data warehouse on a separate SMP cluster or grid computing array. Such non-SQL processing frequently requires large, time-consuming data extractions from the data warehouse – potentially up to several billion rows that have to be exported from the database to the analytics server. This approach works directly against the principal of minimizing data movement – a return to the inefficient world before Netezza. Users and their companies pay the price: in the cost and complexity of a separate analytics server as well as analyses based on stale or incomplete data.

Bringing analytics into the appliance is a natural evolution of the performance and appliance innovations that have made Netezza the success it is today. The same innovative architecture used for executing SQL commands is now enabled for the non-SQL algorithms used by complex analytic applications. Netezza now offers non-SQL programmability directly within the NPS system, typically embedded within the SQL functions of the appliance. Analytic applications run "on stream" – at streaming speeds within the database, with minimal data movement. By moving analytic functions to the SPUs, Netezza's massively parallel streaming processing operates on data where it resides, with unprecedented efficiency and a corresponding leap in performance.

Streaming Analytics™: All the Data, All the Time



Streaming analytics with Netezza provides a distinct advantage in a broad range of applications: for spatial analysis, text mining, risk-profiling, real-time pricing, network monitoring, fraud detection and many others. But in addition, the ability to run a complex analysis against a huge live database, without the delays and costs of moving data to separate hardware, opens up new types of analyses previously out of reach. Here are just a few examples of how the NPS appliance is being used for streaming analytics:

Geospatial Analytics: Geospatial analysis performs operations such as combining multiple maps or map layers according to predefined rules, or identifying regions within a specified distance of one or more features, such as roads or rivers. Geospatial analysis is used for solving problems like: "Find all properties within 10 miles of Hurricane Katrina's eye path" or "Find all properties in Massachusetts that physically straddle county boundaries." A Netezza partner is building an entire geospatial library on the NPS system, with analytics embedded within SQL functions. Users will be able to run geospatial applications on huge, comprehensive data sets, taking advantage of Netezza performance to make rapid and informed decisions.

Predictive Model Scoring: Many companies use predictive modeling to finely segment their customers and make real-time decisions about promotions, pricing, fraud and other applications. This typically involves a time-consuming process: after transaction data is loaded into the data warehouse, the company performs large extracts to a separate cluster server system. The data must be denormalized and then fed back into a predictive modeling application for the actual scoring. Eventually, a score for each customer is loaded back into the data warehouse. The total round-trip can take hours, dominated by the latency of large data transfers. This entire process can now be done within the Netezza appliance, in a fraction of the previous time, providing real-time offers and promotions to the right customers.

"Fingerprinting" with Hashing Algorithms: The Message-Digest algorithm 5 (MD5) is a standard cryptographic hash function with a 128-bit hash value. It is commonly used to store passwords and ensure that files transferred are intact. It is also used in chain of custody document fingerprinting. By performing the hash directly "on stream," the NPS system runs hash algorithms on millions or even billions of records in seconds. This is typically hundreds of times faster than today's method, which requires moving data from the warehouse to a supercomputing grid, performing the hash on the data and then loading the results back to the warehouse.

"Fuzzy Text" Search Analytics: Fuzzy text search analysis uses algorithms that provide a "best guess" of most likely results. One example is the Levenshtein edit distance algorithm, which calculates how many text edits would be required to manipulate, for example, "Madison Avenue" into "Main Street." This type of algorithm is used by many text searching scenarios which require analysis of billions of text records, such as data cleansing of names and addresses for marketing campaigns, or national security applications for complex analysis of names in port of entry data. These types of capabilities open the Netezza appliance to analysis that was not only performance-constrained in the past but simply impossible through a SQL interface.

The Netezza Developer Network

To take advantage of these new capabilities and create widespread industry impact, Netezza has created the Netezza Developer Network (NDN), a community of development partners that recognize the power of bringing analytics to the data. The network provides a dynamic development environment for new algorithms and applications, with unique opportunities for knowledge sharing as well as access to an open development platform for innovation.

Through the Netezza Developer Network, dozens of developers of analytic applications around the world are leveraging Netezza's high-performance analytic platform for competitive advantage. Members of this growing community include leading providers of analytic solutions and services such as SAS, SPSS, Multi-Threaded, Inc., 10e Solutions, IISi, Rate Integration, Carnegie Mellon University and many others. These innovators are developing new algorithms and applications, and leveraging Netezza's streaming architecture as the backbone for a new generation of analytic applications.

Extreme Performance for Analytics

Since Netezza launched its original NPS system, customers worldwide have taken advantage of the power and simplicity of the Netezza approach for a broad range of complex analytic challenges. Now Netezza is "opening the appliance" to customer and partner organizations who want to move to the next level of business analytics, whether for day trading, scoring potential customers, uncovering fraudulent behavior or a myriad of other scenarios. Through the Netezza Developer Network, a highly motivated, highly productive community of developers is helping customers gain analytic insight in ways that were previously impossible.

Just as Netezza transformed data warehouse processing, it is now redefining the world of business analytics. For the first time, users can run complex algorithms against terabytes of fresh data. Critical results are available in seconds or minutes, rather than hours or days. Organizations can take advantage of analyses previously out of reach – now executed on stream, in Netezza's revolutionary analytic appliance.

For more information about "on stream" analytics, please visit www.netezza.com/ndn.

About Netezza

Netezza (NYSE Arca: NZ) is the global leader in analytic appliances that dramatically simplify high-performance analytics for business users across the extended enterprise, delivering significant competitive and operational advantage in today's information-intensive marketplaces. The Netezza Performance Server® (NPS®) family of streaming analytic™ appliances brings appliance simplicity to a broad range of complex data warehouse and analytic challenges. Customers who are realizing the benefits of Netezza appliances include Ahold, Amazon.com, CNET Networks, Debenhams, Department of Veterans Affairs, Epsilon, Neiman Marcus, Orange UK, Premier, Inc., Ross Stores, Ryder System, Inc., The Carphone Warehouse, the US Army and Virgin Media. Based in Framingham, Mass., Netezza has offices in Washington, DC, the United Kingdom and Asia Pacific. **For more information about Netezza, please visit www.netezza.com.**

Netezza Corporation : 200 Crossing Boulevard : Framingham, MA : 01702-4480
+1 508 665 6800 tel : +1 508 665 6811 fax : www.netezza.com

NETEZZA
The Power to Question Everything™