JUKU
Think Global, Act Local!

# Democratizing Big Data Analytics for the Masses

Prepared for DataCore
November 2016

Think Global, Act Local!

## TABLE OF CONTENTS

# EXECUTIVE SUMMARY

Big Data Analytics dramatically impact how we adapt our businesses to capitalize on changing conditions. While declining costs for storage, CPUs and networking make it possible to collect data and store it safely; analyzing the data in real time remains a highly challenging undertaking. Only the most affluent of organizations can invest in the colossal computing resources necessary to keep up with the velocity of inputs. Even they must often resort to renting time on supercomputers or massive server clusters in the cloud, because it simply cannot be done in-house.

*We are quickly moving from traditional applications and databases to next generation scale-out and in-memory tools, causing the size of the infrastructures to grow massively, making Big Data Analytics adoption practically impossible for the vast majority of organizations.*

The anticipated payoffs depend on reaching timely insights from the barrage of data by continuously tweaking queries and models that uncover or forecast new trends. Different departments require simultaneous access to the same data using different tools to determine the influence on them. This leads companies to explore exotic, albeit expensive alternatives. Next generation scale-out and in-memory tools are being tried in place of traditional applications and databases. Unfortunately, these cause the size of infrastructures to grow massively (literally thousands of nodes), making the adoption of Big Data Analytics practically impossible for the vast majority of organizations.

**There are essentially two major issues:**

- On one side, we have an architectural design problem. Web-scale approaches attempt to break the task into small chunks and split data over many machines so they can be worked on in parallel. Even though this may improve computational throughput, spreading the data and merging results from numerous nodes comes with high costs. Scale-out clusters quickly mushroom into wildly expensive infrastructures that incur long inter-node communication delays and programming complexities. What works for traditional business intelligence using long running post-processing batch jobs is simply too slow for real-time analysis.
- It also leads to a second problem – hiring the rare skills to implement and program such distributed systems. This is especially difficult with unfamiliar tools and uncommon programming techniques. Most organizations would prefer to adapt the tools they are accustomed to (like relational databases), but find them to be very I/O constrained, since they are designed for data consistency. New options such as in-memory databases are also prohibitively pricey.

**A breakthrough in Parallel I/O – A fresh and cost-effective alternative**

Recently, a different approach is proving to be very promising. It involves a complete redesign of the once serial I/O process by taking full advantage of modern multi-core CPUs and RAM in parallel. In this way, Big Data Analytics can be tackled without application reprogramming using conventional databases on just a few machines.

It's a game changer resulting in dramatically faster I/O performance for database and analytic applications without changing a line of code or acquiring expensive hardware. Thanks to this layered technology, it is possible to analyze data in real time using familiar tools with far fewer servers, enabling even smaller organizations to benefit from it.

The software is called DataCore Parallel Server, abbreviated here as "DPS". DPS leverages multi-core CPUs and RAM to both parallelize I/O operations and eliminate I/O queuing delays inside the server. It achieves the lowest latency and the highest number of IOPS without costly investments in hardware or software redesign. Notably, DPS enables all applications to benefit from it, and not only those written to work in parallel. It takes full advantage of the computing cores and the high-speed memory bandwidths and internal caching and pathways available within today's multi-core chip sets. DPS concurrently directs I/Os and effectively optimizes the available bandwidth to maximize the use of RAM caches further accelerating response to approximate the speed of in-memory databases for all application workloads.

*With DataCore Parallel Server it is possible to reduce the number of nodes across which applications are spread and substantially reduce inter-node communication as well as I/O latency down to µSecond level without needing proprietary and expensive hardware.*

Some of the time critical, I/O-intensive applications where DPS could be of most value include:

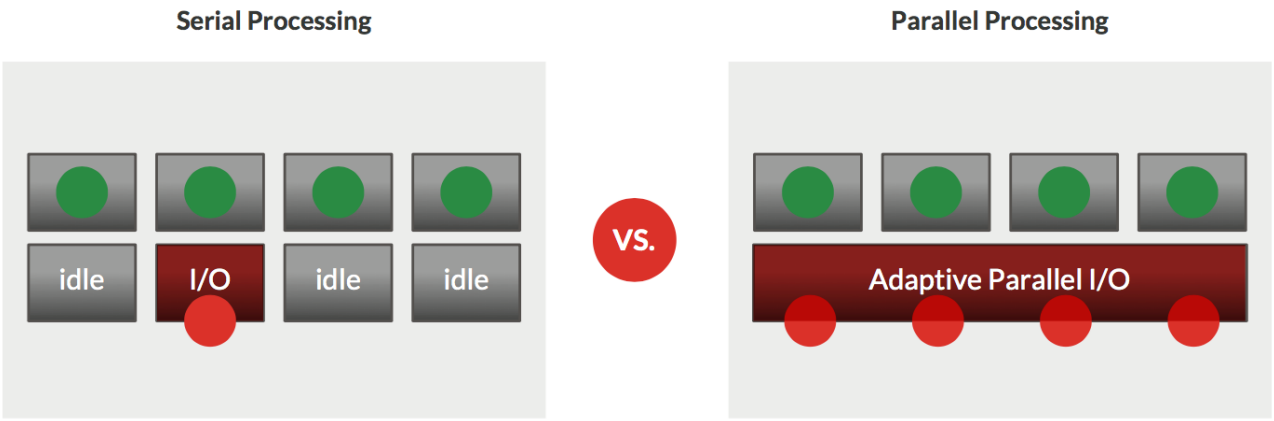
- Large volume / high velocity transactional processing
- Real-time customer offers, ad placements and promotions
- High frequency iterative analytics involving a combination of data in motion and data at rest
- Inventory and resource optimization supporting just-in-time manufacturing and distribution
- Fabrication / construction fault detection
- Fraud alerts and prevention
- Threat assessment (law enforcement, military intelligence and weather forecasting)

# RETHINKING THE I/O PROCESS

Regardless of the commercial operating system or hypervisor being used, and their ability to schedule several concurrent workloads on the same server, they process I/O operations serially. All workload requests are placed in long queues while waiting on the one I/O worker to get to them. Since the applications cannot make progress until the data is read, updated or written, they are forced to go to sleep. Technologies like NVMe reduce the protocol overhead to access storage and networks but, again, they do not address the bottleneck induced by the wait time needed to service and process I/O requests serially and coordinate and manage I/O queues.

*I/O processing even in the most modern kernels is still asynchronous and serialized, forcing applications to sleep while waiting on*

It seems that software engineers have not paid enough attention to how hardware has been evolving. Many years ago, x86 CPUs were fairly simple in design, but scalability remained an issue. On the other hand, powerful scale-up symmetric multi-processing (SMP) systems were very expensive and few knew how to program them. We started seeing more and more single-task servers relying on faster clock speeds, and later, scale-out clusters were chosen to run complex tasks.




Today inexpensive CPUs from off-the-shelf servers offer dozens of cores per socket, and large memory configurations with high bandwidth to memory so they effectively behave like a large SMP system on a chip. DataCore Parallel I/O and caching technologies fully leverage SMP power and memory bandwidths for exceptional speed, essentially collapsing what used to take many physical servers to do, into a single, ultra-fast node. The cores working in parallel effectively multiply the amount of I/O processing that can be performed locally using far fewer servers more efficiently to reach quicker and more timely decisions. Instead of being constantly stalled by network latencies between clustered machines, the bulk of data transfers can now occur in parallel at blazingly fast intra-computer and memory speeds between processes operating on separate cores within the same server. In other words, DataCore enables one to 'do more with less' by collapsing the number of servers needed to get the job done.

Without advanced software to tap on-chip I/O parallelism, much hyped technologies including NVMe and Flash-based storage, have only a marginal impact. They merely shave some time off for the one serial I/O worker who simply can't keep up with multiple simultaneous requests.

*Instead of managing IOs serially and appending them to a queue, Parallel I/O technology uses all the available CPU cores to process them as they come, without delay.*

In contrast, the combination of parallelism and sheer speed, enables DataCore software to sidestep layers of queuing, so multiple threads can transfer data simultaneously to cache without going to sleep. Consequently, Big Data may be analyzed as it arrives without delay.

Thanks to this massive SMP software approach, a far larger number of I/Os can be performed in a shorter time with substantial improvement in the response time. The inherent parallelism of relational databases, for example, can suddenly be realized to its full potential, extending them from OLTP (online transaction processing) workhorses to sophisticated OLAP (online analytical processing) engines.

# BENEFITS OF I/O PARALLELIZATION IN BIG DATA ANALYTICS

Thanks to parallelized I/O processing, applications are able to perform almost all of the reads and writes in microseconds without needing expensive proprietary hardware. A persistent storage layer is still necessary - but now performance is decoupled from capacity, giving the data scientists maximum freedom on hardware choices.

*Thanks to parallelized I/O processing, the application will be able to perform almost all the reads and writes in microseconds and performance is decoupled from capacity giving the end user maximum freedom on hardware choices*

At the same time, because the parallel I/O is transparent to the applications, there is no need to modify code to take advantage of additional resources. This also means that the benefits of this technology are available to everyone - even to the most traditional of applications, like SQL Server, Oracle and SAP, - allowing to race through mammoth datasets. In other cases, by parallelizing I/O operations, the end user can avoid purchasing expensive in-memory database options and save money.

Parallelization of I/O processing has other positive aspects. Since a single server can do much more, it is now possible to consolidate more workloads in fewer nodes, which means:

- Less hardware, meaning also less power consumption and a lower data center footprint
- Fewer software licenses
- A smaller infrastructure, hence, easier to manage

For applications deployed on public clouds, the benefits multiply. DataCore has already demonstrated that a low cost virtual machine (VM) running DPS on Azure or Amazon Web Services (AWS) outperforms the top of the line, premium-priced SSD-based VMs. Translation: rent cheaper resources and for a shorter amount of time. In fact, by substantially collapsing the number of nodes needed, some companies may reconsider bringing the apps back under their control in the privacy of their own data centers.

*No matter if the Big Data infrastructure is deployed on-premises or on the cloud the cost of its implementation is far lower than any traditional option, making it possible to do more with less or more in less time.*

All these aspects combined, make Big Data Analytics accessible to the masses. Whether it is run on-premises or on the cloud, the cost and effort to gain meaningful insights from real-time analysis has now become affordable for organizations both large and small. It is far more affordable than any traditional option, making it possible to do more with less or more in less time.

# BOTTOM LINE

Everyone in the private and public sector is eager to employ Big Data Analytics to glean better understanding from the ecosystem around them in order to make faster and better informed decisions. DataCore Parallel Server software helps you process high velocity transactions and analyze enormous amounts of data in a fraction of the time without resorting to exotic supercomputers, massive server clusters or expensive in-memory code redesigns.

*DataCore Parallel Server software helps you process high velocity transactions and analyze enormous amounts of data in a fraction of the time without resorting to exotic supercomputers, massive server clusters or expensive in-memory code redesigns.*

Its state-of-the-art adaptive Parallel I/O technology has been proven to achieve ultra-fast performance for the most challenging enterprise applications, resulting in better business outcomes with far fewer servers, no programming changes and far less complexity. DPS is transparent to applications whether they work on files or block devices. The microsecond response time and millions of IOPS per server multiply overall computational effectiveness so processing can be performed locally using far fewer servers more efficiently to reach quicker and timelier decisions.

IO parallelization is a game changer in Big Data Analytics. DataCore has the right/appropriate solution to help smaller organizations adopt it, and larger ones get/achieve faster results at a lower cost.

# JUKU

### WHY JUKU

Jukus are Japanese specialized cram schools and our philosophy is the same. Not to replace the traditional information channels, but to help decision makers in their IT environments, to inform and to discuss the technological side that we know better: IT infrastructure virtualization, cloud computing and storage.

Unlike the past, today those who live in the IT environment need to be aware of their surroundings: things are changing rapidly and there is a need to be constantly updated, to learn to adapt quickly and to support important decisions - but how? Through our support, our ideas, the result of our daily global interaction on the web and social networking with vendors, analysts, bloggers, journalists and consultants. But our work doesn't stop there - the comparison and the search are global, but the sharing and application of our ideas must be local and that is where our daily experience, with companies rooted in local areas, becomes essential in providing an honest and productive vision. That's why we have chosen: "think global, act local" as a payoff for Juku.

### AUTHOR

Enrico Signoretti is an analyst, trusted advisor and passionate blogger (not necessarily in that order). He has been immersed in IT environments for over 20 years. His career began with Assembler in the second half of the 80′s before moving on to UNIX platforms until now when he joined the "Cloudland". During these years his job has changed from highly technical roles to management and customer relationship management. In 2012 he founded Juku consulting SRL, a new consultancy and advisory firm deeply focused on supporting end users, vendors and third parties in the development of their IT infrastructure strategies. He keeps a vigil eye on how the market evolves and is constantly on the lookout for new ideas and innovative solutions. You can find Enrico's social profiles here: http://about.me/esignoretti