White Paper

# A New Breed of Servers for Digital Transformation

Sponsored by: IBM

Peter Rutten
January 2017

## IDC OPINION

Digital transformation is not a buzzword. IT has moved from the back office to the front office in nearly every aspect of business operations, driven by what IDC calls the 3rd Platform of compute with mobile, social business, cloud, and big data analytics as the pillars. In this new environment, business leaders are facing the challenge of lifting their organization to new levels of competitive capability, that of digital transformation — leveraging digital technologies together with organizational, operational, and business model innovation to develop new growth strategies. One such challenge is helping the business efficiently reap value from big data and avoid being taken out by a competitor or disruptor that figures out new opportunities from big data analytics before the business does.

From an IT perspective, there is a fairly straightforward sequence of applications that businesses can adopt over time that will help put direction into this journey. IDC outlines this sequence to explain the maturity phases of this journey. This sequence starts in familiar territory such as web and cloud; then continues toward newer terrain, like open source and next-gen apps; and finally it enters the new world of advanced data analytics, predictive analytics, and cognitive analytics. This, in practical terms, is what it means from an IT perspective to digitally transform.

Digital transformation also demands a new approach to data management. Established proprietary relational database management systems (RDBMSs) have served the enterprise well in addressing the requirements of the first two platforms of computing, but the rapidly evolving application technologies of the 3rd Platform, with their demand for managing high volumes of changeable data and development agility, require database technology that is also rapidly evolving and that carries a low cost barrier to adoption. It is for this reason that open source DBMS has become a mainstay of such application development.

The DBMSs themselves vary in design to address the varied needs of 3rd Platform applications: Some are document oriented, some are wide column stores, and some are key-value stores. High-volume data collection and ordering environments, such as Hadoop, are important too. Relational database systems are also part of the mix. The open source nature of these database systems ensures a low cost of entry, and because the database systems are community managed — driven by a steady stream of contributions to the community from application development practitioners — they can evolve more rapidly and nimbly and with a more practical orientation than proprietary products controlled by a vendor-employed development team.

But what is critical to keep in mind is the infrastructure needed to execute this sequence on. All these applications benefit from a robust, high-performing one- or two-socket server infrastructure. Yet when it comes to evaluating one- or two-socket offerings in the marketplace, most businesses only compare

the systems that are available on today's dominant processor architecture. For sure, they believe that they are performing a solid vendor comparison because they are weighing the specifications of various vendors even though these vendors use identical processors for the engine of their products.

What these businesses are overlooking are the advantages that can be gained from a powerful alternative processor, namely POWER8. POWER8-based servers are scalable, have proven to be an easy fit in the datacenter, work well with emerging apps, and can be cost effective. What's more, POWER8 processor technology is at the heart of the OpenPOWER Foundation, in which hundreds of leading technology companies collaborate on developing an ecosystem of innovative, standardized acceleration technologies. Among the accelerators that OpenPOWER members have developed to date are technologies to speed up I/O throughput and processing speed. IBM says that it regards these acceleration technologies from the OpenPOWER partners as the pillar of its value proposition for big data analytics and accelerated computing. IDC believes that such acceleration technologies are becoming a critical part of compute for enterprises.

IDC also believes that not including POWER8-based servers in a comparative evaluation of one- or two-socket offerings is detrimental to the quality and objectivity of that evaluation and may ultimately prevent a business from selecting the most suitable and cost-efficient infrastructure for its digital transformation journey. This white paper takes a closer look at IBM's OpenPOWER LC servers built with the POWER8 processor and with such acceleration technologies as NVIDIA Tesla P100 GPUs, NVIDIA NVLink, and CAPI.

## IN THIS WHITE PAPER

This white paper discusses the digital transformation journey that most organizations are – willingly or unwillingly – on today, why a robust one- or two-socket server infrastructure is a critical component of that journey, and what some of the caveats are for one- or two-socket servers. This white paper defines a road map on one- or two-socket servers for digital transformation in three stages, each with multiple steps – from running simple stateless web applications to adopting open source DBMSs to cloud to predictive modeling. It then talks about the various required parameters for an infrastructure of one- or two-socket servers and goes into the current market landscape for such servers. Finally, this white paper discusses IBM's OpenPOWER LC servers, which run on POWER8, and some of the differentiators of this portfolio. Challenges and opportunities for both IBM and customers are discussed just before the paper's conclusion.

## SITUATION OVERVIEW

### The Digital Transformation Era

Digital transformation is not just a buzzword but the approach by which enterprises drive changes in their business models and ecosystems by leveraging digital competencies. IDC identifies five stages of maturity with regard to the progress businesses have made toward digital transformation (the percentages represent data from IDC's *Digital Transformation MaturityScape Benchmark Survey,* February 2015).

Resisters (14.2%) make up the rear guard, and they provide weak customer experiences and have a defensive posture toward digital. The next category is the Digital Explorers (31.8%) that offer digitally enabled products, services, and experiences albeit inconsistently and not well integrated. The third group are the Digital Players (32.4%) that provide consistent but not truly innovative products, services, and experiences. The fourth segment are the Digital Transformers (13.6%) that are leaders in their markets,

providing innovative products, services, and experiences. And at the front lines are the Digital Disruptors (9%), who are remaking existing markets and creating new ones to their own advantage.

As the data discussed previously indicates, more than 50% of business now fall in the Player, Transformer, or Disruptor categories. The other half are at risk of losing their competitive edge.

## *Why a Robust Infrastructure on Volume Servers Is Critical for Digital Transformation*

The dot-com burst and ensuing recession ushered in an era of reduced IT budgets that led to high demand for cheap hardware running either Windows or Linux. Since those days, data volumes have exploded, the business need to analyze large amounts of data has become critical, and one- or two-socket infrastructure with Linux and open source software has become the most innovative, cost-effective way to engage in today's digital transformation.

But as customers start to engage in this transformation, a question mark looms over the infrastructure they need, not just for the next 12 or 24 months ahead but also 3–5 years out. Businesses that are in the Digital Explorers and Players categories, for example, may be looking to invest in one- or two-socket infrastructure for a few new business applications they are developing and that they do not want to run on their existing scale-up systems. IT at these businesses needs to keep in mind that not all one- or two-socket boxes are created equal and that digital transformation applications will evolve dramatically beyond just the few new apps they intend to launch today.

This white paper presents a road map for digital transformation applications in three stages that demonstrates that a one- or two-socket environment must be extremely robust and high performing while cost effective.

## *One Caveat: The Operational Expenses of Volume Servers*

There are multiple reasons why one- or two-socket servers have been a popular infrastructure choice even before digital transformation became the new mandate. But what has been overlooked is operational expense (opex). Most one- or two-socket infrastructures installed today run at very poor utilization rates, even if they are virtualized. Common virtualization in which the operating system (OS) is not shared has had an impact on opex as every operating system (OS) instance in every virtual machine (VM) needs to be provisioned, life cycled, and managed; backed up; and protected with a disaster recovery strategy. IDC estimates that from 2010 to 2020 server sprawl will cause a quadrupling of server management and administration costs, as well as power, cooling, and datacenter footprint expense increases. In other words, opex may reach a level that is too great a burden for the business and costs much more than anticipated.

To counter runaway opex, IT must consider high-performance one- or two-socket systems with very high utilization rates that require fewer physical systems in the datacenter. Such systems will operate with lower expenses for server management and administration, power and cooling, and datacenter space. They also provide additional advantages such as fewer replication tasks and less complexity for disaster recovery. Ultimately, they enable unburdened employees to focus on revenue-generating and performance-enhancing tasks instead of maintenance operations and persistent workarounds to solve for cascading inefficiencies.

## A Road Map for Digital Transformation on Volume Servers

### *The Basics*

Digital transformation requires that a business redefines how it generates revenue and monetizes products and services, with a strong focus on contextualized and personalized customer experience and business efficiency. It also means centering product and services innovation on customer experience and changing the way the company works to be as agile and dynamic as possible – first by gaining insights and responding to the market with innovative products; then by developing products based on the ability to predict market behavior; and ultimately by creating products that completely reshape the market, possibly in disruptive ways.

In terms of the applications and infrastructure that support this journey, the first step – one that most firms have made at this point – is the adoption of cloud apps.

#### Volume Servers for Stateless Applications

One- or two-socket infrastructure is very well suited for stateless applications such as VDI and many web applications. "State" is the ability of an application to retain such personalized information as to who a user is or what button the user just clicked. *Stateful* applications need a way to keep that information either in a client, the web tier, a database, or in cache. *Stateless* applications do not require such data. Stateless applications are therefore suitable for scaling out as they do not present the challenge of sharing the user's state between the servers. This makes it easy to spin up more servers as needed for such applications.

#### Volume Servers for Web Applications

Traditionally, the biggest concern with web applications has been response time and scalability. End users are demanding instant response times, and for web applications with high traffic volume, this presents a challenge. Scalability helps with improving response time by either adding more servers (one- or two-socket servers) or by adding more – or more powerful – CPUs in a single-node server (scale up). Both approaches are successfully used today.

Another pressing issue is that the way in which application developers' work is shifting to more flexible and agile development styles using a composite architecture. A major development in the industry is the use of microservices – services that are independently created and fine-grained and that can be assembled into application solutions. Infrastructure for developing and running these applications needs to support this modular approach, have a cloud-like structure, and be fully mobile enabled –requirements that a one- or two-socket infrastructure with the right development platform and tools can provide.

#### Volume Servers for Cloud

There continues to be quite a bit of confusion in the marketplace as to what a cloud exactly is, and what the differences are between private, hybrid, and public clouds. Often, IT finds itself confronted with a mandate to "put it in the cloud," typically for perceived cost reasons. To facilitate this discussion that takes place within many organizations, the following few paragraphs discuss on how IDC defines cloud (see *IDC's Worldwide IT Cloud Services Taxonomy, 2015,* IDC #258348, September 2015). Readers can choose to skip this section if they feel sufficiently informed on the subject:

- **Public cloud services** are shared among unrelated enterprises and/or consumers; open to a largely unrestricted universe of potential users; and designed for a market, not a single enterprise.

- **Private cloud services** are shared within a single enterprise or an extended enterprise, with restrictions on access and level of resource dedication, and defined/controlled by the enterprise, beyond the control available in public cloud offerings. In the private cloud services world, there are two major options:
  - **Enterprise private cloud.** In this private cloud scenario, an enterprise typically either acquires a pre-integrated cloud services system or integrates component software and hardware elements and operates the cloud service for its own use. The enterprise sometimes contracts with a third party for integration and/or operational services. An enterprise private cloud may be run in the enterprise's own datacenter or may be colocated in a third-party facility.
  - **Hosted private cloud.** In this private cloud scenario, third-party commercial cloud service providers offer customers access to private cloud services that the service providers have built, own, and operate. Within the hosted private cloud world, IDC identifies two different hosted private cloud deployment models:
    - **Dedicated hosted private cloud,** in which service providers stand up a private cloud system that is fully dedicated to the customer for an extended period of time. This model is essentially a cloud version of traditional managed hosting offerings.
    - **On-demand hosted private cloud,** in which service providers dynamically provision resources for dedicated use from a shared pool — often from the same pool as their public cloud offerings.
- **Hybrid cloud services** are the integration and consolidated management of cloud services with other cloud services and/or noncloud resources (systems, apps, and databases). Hybrid cloud services include "public-public," "public-private," and "private-private" combinations as well "cloud-noncloud" combinations. The inclusion of noncloud resources typically requires front ending the resource with cloud services interfaces (e.g., RESTful APIs).
- **One- or two-socket clouds** are not suitable for every kind of workload, because not every workload is suitable for running on one- or two-socket systems. The cloud is good for enterprise-level apps to roll out and control at a single point. It is not good for mission-critical data sets. Furthermore, clouds built on low-cost commodity architecture cannot provide the high availability (HA) requirements that certain workloads require.

## *The Next Step to Open Source and Next-Gen Applications*

### Volume Servers for Open Source Applications

One- or two-socket infrastructure and open source software go hand in hand because of their combined cost-effectiveness. During the years that one- or two-socket servers became the dominant infrastructure in datacenters around the world, open source software washed across the software landscape like a spring tide. According to Sonatype, which recently released its 2016 State of the Software Supply Chain Report, the number of open source component download requests increased to 31 billion in 2015 from 17 billion in 2014, an 82% increase year over year. Sonatype says that 10,000 new component versions are introduced daily across development ecosystems. Businesses are adopting the open source model across industries for cost reasons, for the ability to engage a vast community of developers that "speak" the same languages, for rapid innovation, and for enabling connections between assets via open APIs that drive new opportunities.

## Volume Servers for Open Source Databases

One particularly valuable resource for modern businesses on a digital transformation journey is the availability of numerous (indeed, hundreds) open source databases, both relational and nonrelational. Before the emergence of one- or two-socket servers as an alternative infrastructure and open source as a software model, the database landscape consisted primarily of Oracle, Microsoft SQL Server, and IBM DB2. These proprietary relational databases have become expensive, charging hefty licensing fees per core. As data volumes rise, and as data is increasingly complex and unstructured, including streams, tweets, video, images, and sensor data, more and more businesses are, partially or entirely, shifting to open source databases.

Such databases are chosen to meet the specific needs of the applications, which often include the ability to accept new data formats without modifying a schema; the ability to scale up or down at will; the ability to handle large, complex objects (documents) in a high-performance manner; and the ability to accept and process data that comes in a wide variety of formats (such as IoT data) or at a great rate of speed (such as streaming data). The established RDBMSs tend to lack the flexibility to meet these needs, and their licensing models constrain their use in a highly scalable manner. Also, the experiences of developers using these open source DBMSs result in intelligent contributions to the code that make the technology highly responsive to changing developer needs. Even open source products that are controlled by vendors (such as MySQL and MongoDB) evolve rapidly because of developer contributions and input.

Being improved on an ongoing basis by a large development community, open source databases such as PostgreSQL have reached enterprise-level quality and are ideal for scaling out because adding more cores to expand them will not ravage an organization's IT budget. Furthermore, nonrelational or NoSQL databases such as MongoDB and Cassandra are increasingly popular for their ability to store any type of unstructured data, and they are designed for deploying and massively scaling new applications. In August 2016, DB-Engines, a monthly updated popularity ranking of databases, listed MongoDB in fourth place after Microsoft SQL Server, PostgreSQL in fifth place, and Cassandra in seventh place after DB2. Redis, the number 1 key-value store in the DB-Engines rankings, was listed in tenth place.

Open source also offers the promise of a low cost of entry, with additional costs corresponding to the actual scale of usage. This contrasts with proprietary software that is offered on the basis of a perpetual use license fee, usually quite large, that is set based on the expected "high water mark" of software usage in terms of "sockets" or "cores." Then, from time to time, the vendor will expect additional fees to be paid as the user increases usage. With open source, you can start by downloading and compiling the community edition of the open source software at no charge. The team becomes familiar with it and starts building applications.

At some point, the enterprise will find it necessary to get a support subscription for the DBMS. This enables the team to have access to tested precompiled binaries and usually includes, in addition to the basic open source components, a set of tools and utilities that provide the kind of data governance, security, and high availability/disaster recovery (HA/DR) support that enterprises expect. Still, these are monthly subscriptions, charged on a pay-as-you-go basis. The actual subscription fee models vary from vendor to vendor but in all cases are designed for maximum user flexibility in usage, which is important given the highly scalable requirements of 3rd Platform databases. These fees are considerably lower than the license and maintenance fees for roughly equivalent proprietary DBMS configurations because so much of the testing, support, and even development work is done for free by advanced developers who contribute code to the open source base, especially in the case of community-managed open source products.

### Volume Servers for Next-Gen Applications

IDC predicts that by 2018, enterprises pursuing digital transformation strategies will more than double software development capabilities and that two-third of their coders will focus on strategic digital transformation applications and services. Next-generation applications, a key component of the digital transformation effort, are distinctly different from traditional applications. They use different programming languages and are designed differently. Next-gen application developers require extremely flexible compute capacity, scalability, infrastructure redundancy, performance, storage capacity, bandwidth, and uptime as well as low-cost per unit of compute, storage, and bandwidth. Some typical characteristics of next-gen apps are:

- They are designed using what are called "microservices," meaning that instead of a monolithic application they consist of a collection of small services that are each responsible for a distinct process, can be deployed independently, and can hook into each other via APIs. A major benefit with regard to scaling an application built with microservices is that a developer only needs to scale the microservices that require scaling, while those that do not can remain unaffected. Monolithic applications need to scale in their entirety.
- They are often stateless and designed with the expectation that the infrastructure they run on is not guaranteed to be resilient. By being stateless, they can simply move to another node in a one- or two-socket environment in case of a failure without the loss of state. If the application itself malfunctions, it simply gets replaced, not fixed.
- They scale based on demand. When demand increases, more instances are spun up automatically. And they are refreshed much more frequently than traditional applications, sometimes several times a day.

## *The Final Step to Cognitive*

### Volume Servers for Cognitive Applications (Machine Learning)

IDC defines cognitive systems as a technology that uses deep natural language processing and machine learning, resulting in understanding as well as the ability to answer questions and provide guidance. The system hypothesizes and formulates possible answers based on available evidence, can be trained through the ingestion of vast amounts of content, and automatically adapts and learns from its mistakes and failures (see *Worldwide Cognitive Software Platforms Forecast, 2015-2019: The Emergence of a New Market,* IDC #258781, September 2015).

IDC predicts that by 2018, more than 50% of developer teams will be embedding cognitive services in their applications (versus about 1% today). Cognitive applications are intensely data driven and have characteristics that are comparable with high-performance computing applications. IDC sees infrastructure for accelerated computing, which is based on one- or two-socket server clusters, increasingly making an entry into clouds and datacenters for cognitive applications.

### Volume Servers for Big Data Analytics

Big data analytics refers to accelerated computing technology to address complex or time-critical big data problems (see *Worldwide Big Data Technology and Services Forecast, 2016-2020*, IDC #US40803116, December 2016). With big data analytics, just as with cognitive, one- or two-socket server clusters are the norm.

The use of large amounts of data started with modeling and simulation by governments, academia, and very large enterprises such as auto manufacturers, consumer goods producers, and national retailers. The emergence of big data was driven by the internet and by mobile and social consumption behaviors.

Today, simulation and analytics have begun to converge in the commercial sector as enterprises are starting to use accelerated computing to solve big data analytics problems that are critical for their competitiveness.

### Volume Servers for Advanced and Predictive Analytics

Advanced and predictive analytics (APA) is the ability to mine structured/unstructured data and develop insights to drive business decisions by using statistical analysis software that depends on large volumes of data and is therefore often used in conjunction with a data warehouse in a one- or two-socket environment.

APA uses a range of techniques to create, test, and execute statistical models. Some techniques used are machine learning, regression, neural networks, rule induction, and clustering. Advanced and predictive analytics are used to discover relationships in data and make predictions that are hidden, not apparent, or too complex to be extracted using query, reporting, and multidimensional analysis software (see *Advanced and Predictive Analytics Software: Market Segments and User Types,* IDC #APA51X, June 2015).

APA used to require expert skills but is increasingly used by business users who are not necessarily experts but who use APA results to make informed data-driven business decisions. APA consists of several software types: development languages for writing models and programming routines; data science productivity enhancement software that helps scientists with packaged algorithms, data preparation operations, and machine learning solutions; business analyst usability enhancement tools that simplify the usage of APA for business analysts by hiding the analytical models; APA applications for LOB users, such as predictive maintenance and banking fraud detection; and embedded APA in applications (e.g., incorporated in call center software to predict churn probability).

## *Required Parameters for Infrastructure on Volume Servers*

For IT to traverse the one- or two-socket server road map effectively, all the way from stateless web applications to cognitive and predictive workloads, the server hardware that the infrastructure is designed with needs to meet a variety of requirements. The road map illustrates that as these workloads evolve, they cannot be successfully deployed on low-end commodity hardware. The demands on the infrastructure will be too great.

The servers need to be easy to configure and support rapid deployment. The servers must also support the next-gen applications that will be the engines of the organization's digital transformation. The servers need to be easy to manage and have to perform extremely well with data-intense workloads. Indeed, they should ideally support various open and standardized acceleration technologies to overcome the diminishing returns of Moore's law. They need to be secure. And they should provide high availability across the stack because downtime is not an option.

One- or two-socket server infrastructure for these workloads must also be manageable without requiring unusual skill set. It is important that they support a broad software ecosystem and are fully cloud enabled. The servers must run on a widely adopted OS for digital transformation and be able to run the full breadth of open source software solutions. Finally, they need be available at a low capex and opex, and they should be easily replaceable when their next-generation successors arrive.

## The Current Market for Linux-Based Volume Servers

The current market for one- and two-socket servers running Linux, is growing rapidly. IDC forecasts that from 2016 to 2020, the number of shipments of such systems will increase with a 7.5% CAGR. In 2020, the Linux share of the one- and two-socket server market will have grown to 42.6% of the total market in terms of vendor revenue.

The processor choice is arguably the single-most-important component of the decision for one- or two-socket servers that CIOs, CTOs, CMOs, and CMTOs need to make. It is therefore important for customers to not only evaluate systems with x86 processors but also include the alternative POWER processor in their comparative research. Given the fact that with POWER, there is a new breed of low-capex and low-opex one- or two-socket offerings on the market with little endian Linux as the OS that seems very well suited for the data-intensive applications required to support the digital transformation – buyers would benefit from comparing these solutions with the x86 products they are considering.

## IBM'S OPENPOWER LC SERVERS: A NEW BREED OF VOLUME SERVERS

Many businesses that are looking to implement digital transformation applications and that do not want to run them on their scale-up environment are investigating one- or two-socket infrastructure. What many of them are not doing is comparing the one- or two-socket offerings on x86 with one- or two-socket IBM POWER systems, the only alternative to x86, which given their performance and price can objectively be considered a price performance leader.

From a performance perspective, IBM POWER's one- or two-socket server portfolio is based on the same IBM POWER8 processor design as its scale-up products, which are by far market leaders in the 8-plus socket segment. IBM POWER8 has outstanding per-core capabilities because of its 8-way Simultaneous Multi-Threading (SMT) versus 2-way Hyper-Threading in x86, high-bandwidth I/O subsystem, and superior memory bandwidth. These same capabilities that drive the market leadership of the scale-up IBM POWER systems have been incorporated into IBM’s OpenPOWER LC line.

IBM POWER servers have been built around the concept that businesses are first and foremost – and increasingly – running business-critical applications for performing data operations. That includes next-generation applications. A typical mobile application, for example, is very data oriented, even if it is a different type of data than rows and columns. The IBM POWER architecture is designed to adapt to new forms of data, whether it is structured data, measured in petabytes, blobs of unstructured data, or streams. The system has very high ingest rates for data, and its caches have been designed to keep the processor busy. By contrast, x86 processors are aimed at a wide variety of purposes, from playing video games to processing spreadsheets to running an application. By focusing on business-oriented applications, IBM has the ability to optimize IBM POWER for those workloads.

When we look at this from the perspective of an application developer, these characteristics allow them to develop differentiated apps that leverage the performance advantages of IBM POWER. Threadiness, for example, means that an application developer isn't just coding for parallel processors (as on x86) but for eight parallel threads within the parallel processors, allowing for a much higher number of simultaneous processes to optimize the app. Existing apps that were running on x86 can, according to IBM, in 95% of cases simply be moved over and will run out of the box. With some tuning and leveraging the caches and the threading, these too will start to perform better.

But developers can realize even greater differentiation with acceleration on IBM POWER. Specifically, IBM's OpenPOWER LC portfolio can be ordered with CAPI, which stands for Coherent Accelerator Processor Interface, an interface between the processor and the I/O that performs much better than a standard PCIe Interconnect, opening the POWER8 processor to coherent, accelerated offload for networking, storage, and compute workloads. IBM's OpenPOWER LC portfolio can also be ordered with GPUs. IBM's LC line has a strong accelerator portfolio including both CAPI and NVIDIA NVLink Technology — for the S822LC — that developers can leverage for innovative apps.

## The Portfolio: From IaaS to Volume Servers to Enterprise Class

The IBM POWER8 portfolio runs from infrastructure-as-a-service (IaaS) offerings to a range of one- or two-socket products to midrange (four sockets) to enterprise level (eight sockets). IBM POWER8 is available as IaaS in SoftLayer, providing POWER with the efficiencies of the cloud. PurePower Systems is IBM's converged infrastructure for cloud. And IBM offers prebuilt cloud implementations to its customers that include hardware, software, and services.

In the one- or two-socket category, IBM offers a portfolio of very high-performing single-socket and dual-socket systems that start at a purchasing price just under $5,000. The one- or two-socket line for mission-critical workloads consists of Power S822, Power S814, and Power S824 (the second number indicates number of sockets). The one- or two-socket servers on Linux line includes the following systems (L stands for Linux, C stands for Cluster) branded as: Power S822LC for Big Data, Power S822LC for High Performance Computing (with NVIDIA NVLink Technology), Power S812LC, Power S821LC, Power S822LC for Commercial Computing, Power S812L, Power S822L, and Power S824L. With this line, IBM has pumped new blood into the volume server segment and is offering a distinct alternative to x86-based volume servers, if not a series of advantages from a value perspective. Its three enterprise servers are Power E850, Power E870, Power E870C, Power E880, and Power E880C.

The portfolio widens even further when looking at IBM POWER systems built by other server manufacturers that are part of the OpenPOWER Foundation. The Foundation licenses IBM POWER technology to third parties and brings other technology vendors that provide innovations for IBM POWER acceleration under one umbrella. POWER servers are currently available or being built by Tyan, Inspur, Supermicro, Inventec, Wistron, Cirrascale, ChaungHe, and Zoom Netcom. Rackspace is expected to start running Barreleye, its bare-metal IBM POWER8 cloud server, any day now in its cloud, and Rackspace and Google are both working on Zaius, which is POWER9 based.

## POWER Volume Server Differentiation

As mentioned previously, one- or two-socket servers for digital transformation demand certain infrastructure parameters. This section investigates how IBM POWER's one- or two-socket server portfolio performs on those parameters, except for the system's performance, which has been discussed previously.

### *Deployment Speed*

Deployment speed means two things: the deployment of the hardware and operating environment and the deployment of applications. For hardware deployment, IBM's OpenPOWER LC systems support the same kinds of, or even the same, tools for deployment as x86 and at comparable speeds. IBM provides POWER VC, which stands for Virtualization Center. This is a tool built on OpenStack for virtualization management and cloud deployments. Its purpose is to automate and simplify the management of VMs on POWER one- or two-socket servers with an easy-to-use user interface and to optimize resource allocation. POWER one- or two-socket servers can also take advantage of such open source tools as JuJu charms for rapid deployment. In addition, IBM has a program called Rapid

Build Solution for boosting deployment speeds by delivering preconfigured and preloaded systems that are ready for customers to plug in and start using. With regard to application deployment speed, when an application setup requires a large amount of processing, POWER's architectural benefits will kick in.

## Next-Gen Application Support

Next-gen application developers benefit from the technical capabilities of a POWER8 one- or two-socket system. ISVs and application developers have reported that a pleasant circumstance of coding on POWER8 is that it doesn't require anything new. They can pick the Linux they like, work with the tools they prefer, and use the coding techniques that they've always been using. In other words, a developer can simply start developing for POWER without having to learn new skills.

Beyond the similarities between coding on POWER8 and other systems, there are distinct benefits that POWER8 provides, allowing application developers to optimize on the architecture and deliver more to their clients using the same code base. For example, developing apps on Linux and IBM's OpenPOWER LC provides benefits of flexible scaling, mission-critical resiliency and reliability, and performance. Among those benefits is also the previously mentioned 8-way threading. It is generally understood that the biggest leap for an application developer is changing from single thread to multithread coding. Going from two to, for example, eight threads is seen as incremental work to parallelize the processes; it is not the same big learning step. What's more, application developers who have been happy writing for single processes on x86 can still write single-threaded processes on POWER8, which will run faster. The native strengths of POWER8 — memory bandwidth, cache structure, and thread density — mean developers can take the same code built on x86 and run it on POWER8 with few or no changes and see improvements. For example, for memory-intensive codes, developers will get a performance boost when moving to POWER8.

## High Availability

A question that is on many people's minds is, whether IBM would actually provide comparable HA features on its one- or two-socket server line as on the enterprise-class scale-up systems, with which it has built its reputation for reliability. What differentiates POWER8 with regard to HA resides in its architecture and has been carried over from its enterprise-class systems to its one- or two-socket line.

Memory buffers that are built into the system, for example, help eliminate soft errors. Soft errors can occur at a chip or system level, are caused by particles or noise, and will alter an instruction in a program or a data value. A soft error will typically bring down lesser architecture and require a reboot. POWER8 also has automatic recover processes designed to recover from internally detected faults. In other words, the system doesn't surface the fault and then ask a software package to deal with it, it fixes it. POWER8 features intelligent memory controllers with replay buffers and error detection so that it knows when there is a problem, which it can then correct on the bus or between the controller and the DIMM. It also features spare DRAM modules.

What IBM has essentially done is built an HA continuum in its one- or two-socket line. IBM’s OpenPOWER LC product line has fewer of these built-in HA features because it is designed and cost optimized for clusters in a cloud deployment. In a cluster, if a node fails another node will take over and processing continues as before. The other end of the one- or two-socket line, such as the S824 or the S822, has more of these RAS features for mission-critical workloads because those products are often deployed as a single system. Within that space, IBM has incorporated as many enterprise learnings as possible. In the higher end, this includes redundant components, but in IBM’s

OpenPOWER LC line, the focus is on design points to remain cost competitive: a better part, a better vendor, and a better process to manufacture.

## Software Ecosystem

The software ecosystem for the IBM POWER8 one- or two-socket line is essentially as broad and varied as the software ecosystem for all Linux. If it runs on Linux on x86, it runs on Linux on POWER, is for the most part a true statement, with one caveat: if an application takes advantage of a proprietary extension, regardless of the type of hardware, then a customer or ISV will have to figure out how to handle that part for POWER8. This is not a unique situation for POWER; the same will occur on x86 when moving software that uses a proprietary extension from one hardware manufacturer to another. Other than that, IBM is confident that code can simply be brought across to POWER8. If it's a compiled code, it just needs to be recompiled; if it's scripted, it simply runs.

According to IBM, when Canonical moved Ubuntu onto POWER, all the scripted code ported over and ran. And of all the company's compiled code, 95% ran. Ubuntu simply recompiled the code, and it ran without errors. The 5% that did not work was due to the use of proprietary extensions and because of a few unique coding tricks that had to be reworked to have the codes run on POWER8.

## Cloud Readiness

IBM's OpenPOWER LC line is designed and cost optimized for cloud deployment. The per-core performance and VM density with POWER8 are a significant benefit for a cloud environment as combined they address a major limitation that datacenters face, which is the cost of floor space. Per-core performance and VM density facilitate consolidation of hardware. IBM says that because of POWER8's VM density, datacenters can run substantially more VMs in the same footprint. And further density advantages can be achieved with containerization, on Docker for example. For MSPs POWER8's VM density advantage means, fairly simply, more revenue per square foot; for datacenters it means lower opex.

When customers deploy applications in the cloud, they need them to be available. The POWER8 HA features discussed previously, such as nonvolatile memory and memory protection, ensure that a cloud on POWER8 one- or two-socket servers is intrinsically reliable. With regard to skill sets, there are no differences — an admin who can run a cloud on x86 can also run a cloud on POWER8, using the same tools.

Finally, one major advantage of POWER in the cloud is that the architecture is completely open and licensable — from the hardware to the code, essentially like the ARM model. Hyperscale cloud providers such as Google and Rackspace have stated that they are developing and manufacturing POWER architecture components for their datacenters, which they've licensed from IBM via the OpenPOWER Foundation. IBM strongly believes that the OpenPOWER Foundation, which allows partners to license every bit of POWER technology, is critical to the long-term success of cloud.

## Low Cost

The general perception is that POWER is always more expensive; however, today anyone can go to IBM's website and purchase a Power System S812LC for $4,800. IDC has discussed this price point, and the price continuum of the entire one- or two-socket line, with IT staff at several organizations and has witnessed firsthand that it changes their perception. IBM recently announced new additions to the LC portfolio that will improve upon the price point even further. While IDC does not compare systems on price or price performance, the metrics for POWER8 suggest that IT buyers would do themselves and their

organizations a disfavor by not performing those price/performance comparisons as part of their one- or two-socket infrastructure evaluations.

## CHALLENGES/OPPORTUNITIES

### For Organizations

#### *Challenge*

For businesses, the greatest challenge is to find the right one- or two-socket server strategy that will support their digital transformation journey. There are many variables: What kinds of applications are we running today? What types of applications will we be running in 12 or 24 months? What are our HA requirements? Do we want to deploy in a cloud? Private, hybrid, or public? What is the right infrastructure?

#### *Opportunity*

The greatest opportunity for these business is to gain significant competitive advantage by making the right infrastructure decision. This can only be accomplished by not only comparing various vendors of x86 infrastructure but also by including OpenPOWER-based infrastructure in the evaluation. Current one- or two-socket products that run on POWER8 are highly competitive, if not advantageous from a price/performance perspective.

### For IBM

#### *Challenge*

IBM Power Systems is well known for its enterprise portfolio of scale-up servers with high-performance characteristics. These systems traditionally ran on Unix but are now also available on Linux, which the marketplace is starting to acknowledge. However, what the marketplace seems to be unaware of is that IBM markets a strong and diverse portfolio of one- or two-socket servers for cloud and for digital transformation applications. These single- and dual-socket systems start at a surprising price point of $4,800. Yet these systems feature the same POWER8 chip that serves as the engine of the large enterprise systems. IBM has to make a publicity splash with these systems, proof to the marketplace that they are higher-performance, more cost effective, and just as easy to use as one- or two-socket servers on x86, and push the market to a point where any business' comparative one- or two-socket systems evaluation includes POWER offerings.

#### *Opportunity*

For IBM, the opportunity is significant. Today, the market for single- and dual-socket servers is essentially in a monopolistic state from a processor architecture perspective. While there is plenty of competition between vendors that market systems with the same processor architecture, this ultimately does not represent true choice for end users. If IBM can show the market that its alternative one- or two-socket offerings on POWER are competitive or even advantageous from a price/performance perspective, there will be a lot of market share to be gained.

## CONCLUSION

Digital transformation is about business model and product innovation leveraging new digital and increasingly mobile capabilities to create experiences that delight customers and satisfy their evolving expectations. The "digital capabilities" included in this description are, of course, a prerequisite for

taking an organization through that journey. They consist of many things, from app developers who can code using microservices to applications that can predict customer behavior to processors than can parallel process instruction sets.

IDC believes that evolving digital capabilities in a business means adopting the right applications running on the right infrastructure. We also believe that there's a staged approach to doing so that will make it easier and more efficient to ultimately get to today's holy grail of cognitive analytics, high-performance data analytics, and advanced and predictive modeling for distinct competitive advantage. This white paper describes those stages in detail.

We also describe what we believe would be required from the infrastructure that serves as the foundation of digital transformation. It can most certainly benefit from being Linux on one- or two-socket servers, although many of the prescribed stages can also be executed on Linux-running scale up, given sufficient utilization, virtualization, and scalability. However, one- or two-socket servers have distinct benefits, as explained previously.

IT leaders need to be very careful, however, with regard to the one- or two-socket infrastructure they choose to build their transformation on. Most one- and two-socket servers that Linux environments are architected with come with only one processor flavor, x86, even if they are available from many different brand-name vendors. IDC believes that businesses that evaluate only one- or two-socket server products that run Linux on x86 are overlooking an opportunity to compare its metrics with IBM's OpenPOWER LC servers. While IDC does not engage in direct comparisons between processors, we do believe that customers should do so thoroughly before they invest in their IT infrastructure for the next five critical years in this era of tumultuous change.

## About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

## Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com

---