



White Paper

A New Breed of Scale-Out Solutions for Digital Transformation

Sponsored by: IBM

Peter Rutten
September 2016

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 and 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.

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.

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 scale-out infrastructure. Yet when it comes to evaluating scale-out 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, more specifically IBM Power Systems. OpenPOWER-based servers built with POWER8 technology 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, numerous accelerator technologies that OpenPOWER members have developed further speed up I/O throughput and processing speed. IDC believes that not including OpenPOWER-based servers in a comparative evaluation of scale-out 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.

IN THIS WHITE PAPER

This white paper discusses the digital transformation journey that most organizations are – willingly or unwillingly – on today, why a robust scale-out infrastructure is a critical component of that journey, and what some of the caveats are for scale out. This white paper defines a scaling-out road map for digital transformation in three stages, each with multiple steps – from running simple stateless web applications on scale-out infrastructure to cloud to predictive modeling. It then talks about the various required parameters for a scale-out infrastructure and goes into the current market landscape for scale-out servers. Finally, this white paper discusses IBM's new breed of scale-out servers on POWER8 and some of the differentiators of IBM's POWER8 scale-out 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 MaturityScope 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 Scale-Out Infrastructure Is Critical for Digital Transformation

To scale out means to build grids of server and storage hardware that are massively parallel and linearly scalable and that have built-in redundancy so that if one unit fails, another takes over. Scaling out on a large scale began with high-performance computing (HPC) led by Google and Amazon in the early 2000s, then caught on like wildfire among datacenters worldwide.

Data volumes were increasing, and the older Unix-based scale-up systems and mainframes were considered too expensive, too complex, and unable to run the emerging wealth of open source software (they have dramatically caught up since then, though). 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 scale-out 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 scale-out 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 scale-out boxes are created equal and that digital transformation applications will evolve dramatically beyond just the few new apps they intend to launch today.

Investing in low-end commodity hardware today will almost certainly lead to limitations on an organization's digital transformation journey. It is advisable to think of a scale-out investment as augmenting an existing scale-up environment for a specific set of current and future business applications that is required to support digital transformation. Each of the applications that together make up this digital transformation journey is intensely data driven. These applications are database driven or streaming data driven, and they have to process ever larger amounts of data in ever more complex ways fast in order for the organization to gain competitive advantage from deep insights.

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

One Caveat: The Operational Expenses of Scale Out

There are multiple reasons why scale out has been a popular infrastructure choice even before digital transformation became the new mandate. Unlimited scalability is one. The ability to rapidly add new workloads by adding a few boxes is another. Creating private or hybrid clouds and seamlessly connecting to public clouds is an important aspect. The flexibility of scale out versus the more complex nature of scale up plays a role and of course the smaller financial commitment in terms of capital expenses (capex) that scale out represents given at a time of rapid technology advancement.

But what has been overlooked is operational expense (opex). Most scale-out 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 cost much more than anticipated.

To counter runaway opex, IT must consider high-performance scale-out 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.

A Scaling-Out Road Map for Digital Transformation

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.

Scale Out for Stateless Applications

Scale-out 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.

Scale Out 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 (scale out) or by adding more – or more powerful – CPUs in a single-node server (scale up). Both approaches are successfully used today.

Scaling out can be advantageous for web applications as the hardware can be allocated in smaller increments. While web applications used to be considered less mission-critical than a business' core data processing workloads, this too is changing, and web applications on a scale-out system need to be protected with load balancing, virtualization, and/or clustering solutions.

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 scale-out infrastructure with the right development platform and tools can provide.

Scale Out for Cloud

The cloud is about economies of scale and a private or hybrid cloud that has been designed on scale-out infrastructure can be highly advantageous if built with robust cloud management tools for monitoring, maintenance, high availability (HA), and disaster recovery (DR). Without such tools, the benefits of scale out can evaporate quickly. A typical cloud consists of a large number of identical low-cost servers (in hyperscale centers they are often density optimized, meaning 0.5 unit per socket and no redundant hardware) that are networked as a grid. By dynamically allocating requested resources to an end user and charging a per-use fee, clouds achieve very high efficiencies – higher than, for example, nonvirtualized servers, virtualized servers, or distributed environments that are not managed as a cloud. This holds true for private, hybrid, and public clouds.

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 (HPC).** 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 HPC world, IDC identifies two different HPC deployment models:
 - **Dedicated HPC**, 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 HPC**, 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).
- **Scale-out clouds** are not suitable for every kind of workload, because not every workload is suitable for running on scale out. 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 HA requirements that certain workloads require.

The Next Step to Open Source and Next-Gen Applications

Scale Out on Linux

Scale out and Linux were born in the same technology era – the early 2000s. Scale out on commodity x86 hardware was considered cheaper in terms of lower up-front costs because it saves on capex, and open source OS Linux was either free or, as a commercial distribution, much more affordable than Unix. Since then, Linux has had an unprecedented run in terms of capabilities and market share. It is now the operating system for cloud, open source applications, open source databases, and HPC.

All hyperscale clouds run Linux, and Linux no longer runs only on x86 – it is now available on Linux-only mainframes and on IBM POWER8 scale-up and scale-out servers. IDC forecasts that 24% of software worldwide will run on Linux by 2020 compared with 16% in 2016. In the server market, Linux's share of worldwide server vendor revenue had grown to 31.2% in 2015.

Linux has been so successful that SAP has decided to run HANA, the database that it has essentially bet the company on, on Linux only. With SAP HANA, companies can now run powerful analytics systems and transactional systems using SAP HANA as an in-memory database deployed on Linux servers.

Scale Out for Open Source Applications

Scale-out infrastructure and open source software go hand in hand because of their combined cost-effectiveness. During the years that scale out 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.

Scale Out 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 scale out 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. These are perhaps a bit less user-friendly, but they have similar functionality and, increasingly, comparable availability and security.

Being improved on an ongoing basis by a large development community, open source databases such as MySQL and Postgres 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. DB-Engines, a monthly updated popularity ranking of databases listed MySQL in second place after Oracle, MongoDB in fourth place after Microsoft SQL Server, and Cassandra in seventh place after DB2 in August 2016.

Scale Out 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 scale-out 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

Scale Out 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 HPC-like infrastructure, which is based on scale-out clusters, increasingly making an entry into clouds and datacenters for cognitive applications.

Scale Out for High-Performance Data Analytics

High-performance data analysis (HPDA) refers to high-performance computing technology to address very complex or time-critical big data problems – including jobs that enterprise server technology alone can't handle (see *IDC FutureScape: Worldwide High-Performance Data Analysis 2016 Predictions – Science and Industry Turn to HPC for Mission-Critical Big Data Challenges*, IDC #259921, November 2016). HPDA is, just as cognitive, springing forth from HPC where scale-out clusters are the norm.

The use of large amounts of data originated in the HPC community where it was used for 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 HPC to solve big data analytics problems that are critical for their competitiveness.

Scale Out 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 scale-out environment. Note, however, that predictive modeling such as SPSS can also be applied to transactions on a scale-up environment (e.g., credit card fraud detection in a banking application that runs on a mainframe).

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 Scale-Out Infrastructure Parameters

For IT to traverse the scale-out 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 scale-out 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-intensive workloads. They need to be secure. And they should provide high availability (HA) across the stack because downtime is not an option.

Scale-out 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 to 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 Scale-Out Servers

The current market for Linux-based scale-out infrastructure, meaning 1- and 2-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 1- and 2-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 a scale-out decision that CIOs, CTOs, CMOs, and CMTOs in businesses around the world 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 scale-out 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 POWER: A NEW BREED OF SCALE-OUT SOLUTIONS

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 scale-out infrastructure. What many of them are not doing is comparing the scale-out offerings on x86 with scale-out IBM POWER, the only alternative to x86, which given its performance and price can objectively be considered a price performance leader.

From a performance perspective, IBM POWER's scale-out 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 the scale-out IBM POWER offerings.

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, the LC portfolio (812LC and 822LC) 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, as well as with GPUs. The LC line and some products in the L line have a strong accelerator portfolio that developers can leverage for innovative apps.

The Portfolio: From IaaS to Scale Out to Enterprise

The IBM POWER8 portfolio runs from infrastructure-as-a-service (IaaS) offerings to a range of scale-out products (1- or 2-socket systems) to midrange (4 sockets) to enterprise level (8 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 scale-out 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 scale-out line for mission-critical workloads consists of Power S822, Power S814, and Power S824 (the second number indicates number of sockets). The scale out on Linux line includes six systems (L stands for Linux, C stands for Cluster): Power S812LC, Power S822LC for commercial computing, Power S822LC for high-performance computing, Power S812L, Power S822L, and Power S824L (with NVIDIA technology). 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 Scale-Out Differentiation

As mentioned previously, scale out for digital transformation demands certain infrastructure parameters. This section investigates how IBM POWER's scale-out 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 POWER scale-up and scale-out 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 scale-out servers with an easy-to-use user interface and to optimize resource allocation. POWER scale-out 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.

Core Applications Support

With regard to support for a business's core applications, like data management, on the IBM POWER scale-out line, all IBM POWER systems whether they are the smallest or the largest in the portfolio have the same integrated operating environment and integrated capabilities. A customer can run the same application at the very lowest and the very highest end of the spectrum, regardless of what that application is. For example, if in the past a customer ran its database on a very large POWER5 machine for reasons of sheer horse power, today, that same database will run on a scale-out POWER8 server with, perhaps 4 instead of 16 sockets, thanks to the POWER8 performance metrics. In short, POWER supports those applications across the entire line.

This is also true for larger x86 systems. A data-intensive application, for example, that runs on a 4-socket x86 system will easily migrate to a 1- or 2-socket POWER system, thanks to the POWER system's workload density. SAP HANA, for example, which has been running on x86 for several years, became available on POWER in 2015. HANA on POWER has seen a significant adoption rate in datacenters, as IBM reported in its recent quarterly update. IDC has spoken with HANA customers that switched from x86 to POWER because they were running into limitations with their x86 HANA appliances. They consolidated HANA and various SAP applications onto a single POWER scale-out system.

Next-Gen Application Support

Next-gen application developers benefit from the technical capabilities of a POWER8 scale-out 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 scale-out POWER 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.

High Availability

A question that is on many people's minds is, whether IBM would actually provide comparable HA features on its scale-out 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 scale-out 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.

IBM says that many of its customers are buying POWER because of nonvolatile memory (NVM) and the ability to correct memory failures before they happen. In particular, customers that were running HANA on x86 and switched to HANA on POWER have mentioned that they made the switch because their previous systems failed due to memory failures.

What IBM has essentially done is built an HA continuum in its scale-out line. The low-end 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 scale-out 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 the 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 scale-out 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

The POWER8 scale-out 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 scale out 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 scale-out 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 scale-out infrastructure evaluations.

Adding Enterprise-Class Scale-Up POWER to a Scale-Out Environment

Finally, customers that are considering POWER8 for scale out should take a look at the Power Systems scale-up line as well. Running both will provide them with architectural advantages, as well as the benefits of years of technology development that is embedded in both the scale-out and scale-up portfolio. POWER-based systems are fully enabled to run open source workloads, with little endian Linux available across the portfolio from the scale-out to the scale-up systems. POWER's Integrated Facility for Linux processors (IFL) allows customers to very economically run even the smallest Linux workloads, such as an emerging next-gen application.

Customers can start next-gen applications on smaller scale-out systems and then migrate them to the enterprise-class POWER system as they begin to scale and start requiring the same enterprise characteristics as the traditional systems of record applications. For example, POWER customers are increasingly running Linux-based open source databases such as MongoDB, MariaDB, and Postgres next to or even instead of their traditional systems of record applications.

IBM Power Systems offers three scale-up systems that are ideal for a digital transformation journey: the E850, the E870, and the E880. They are built around the new and very fast IBM POWER8 processor and come with up to 16TB of memory in the E870 and 32TB in the E880. The systems have strengthened cores and larger caches compared with previous POWER generations as well as two integrated memory controllers with improved latency and greater bandwidth. All three systems have an integrated I/O subsystem and hot-pluggable PCIe Gen3 I/O slots. Various improvements have increased the systems' already well-known reliability, serviceability, and availability, including redundant power supplies, advanced memory protection, and redundant service processors.

For more on the scale up for digital transformation, see *Digital Transformation Decisions: Bringing Your Enterprise-Class Scale-Up Solutions into the Next-Gen World* (IDC white paper).

CHALLENGES/OPPORTUNITIES

For Organizations

Challenge

For businesses, the greatest challenge is to find the right scale-out 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 scale-out 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 scale-out 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 scale out on x86, and push the market to a point where any business' comparative scale-out 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 scale-out 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 scale out, although many of the prescribed stages can also be executed on Linux-running scale up, given sufficient utilization, virtualization, and scalability. However, scale out has distinct benefits, as explained previously.

IT leaders need to be very careful, however, with regard to the scale-out infrastructure they choose to build their transformation on. Most 1- and 2-socket servers that Linux scale-out 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 scale-out server products that run Linux on x86 are overlooking an opportunity to compare its metrics with OpenPOWER-based scale-out servers, specifically IBM's OpenPOWER LC. 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

Copyright Notice

External Publication of IDC Information and Data – Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2016 IDC. Reproduction without written permission is completely forbidden.

