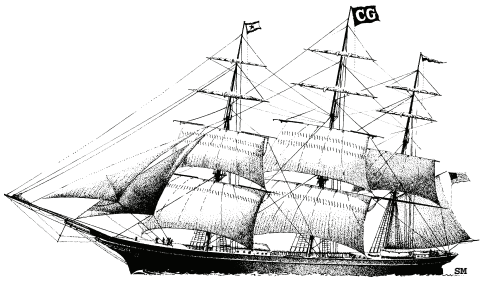
THE CLIPPER GROUP
Navigator™



Navigating Information Technology Horizons℠

Published Since 1993 Report #TCG2007069 June 12, 2007

UNIX Consolidation and Virtualization — IBM Supercharges System p with POWER6

Analyst: David Reine

Management Summary

Owning a car in an urban area is becoming impractical, not only to the individual but to the environment, as well. Owning two, is cost prohibitive. If you only drive your vehicle to commute to work, you use it perhaps 30 – 60 minutes a day, 4% of every day. **Would you tolerate an employee who worked only 4% of the time, or even 12%**? The cost of running a car involves more than just the purchase price. Many city residences do not come with parking; you have to pay extra, often hundreds of dollars per month. You also pay a sales tax and annual charges for registration, maintenance, and insurance. Insurance in many cities is higher than the rate for the same vehicle in a rural setting. The total cost of ownership for an automobile could be twice the purchase price, or higher, not to mention the costs of air pollution. What is the alternative? Cities usually have a reliable public transportation system that will get you to work and back home for a few dollars a day. However, what do you do if you need to go shopping during the week or to the beach on Sunday?

Zip Cars have appeared to satisfy the occasional driving needs of the city resident. Their business model allows you to *share access* to a car with others, for an hour or a day. You can select the right car for your needs, on demand, no matter what your errand: a small car for weekly shopping, a van to bring home a plasma TV. You do not buy a car; you do not pay for parking or insurance. You simply enroll in the Zipcar program, paying an application fee, and you have access to a fleet of cars, paying by the hour, when and where you need one. You have *virtual* ownership of transportation, a sedan to go shopping, or a convertible for the beach. Moreover, 20 of your neighbors can be sharing the same car at different times during the week, *consolidating* all of their driving needs in the same car. This is also the ideal solution for a campus full of students with an occasional driving need, but without the means to own their own car.

A similar situation exists in the data center of every enterprise, large and small. Server farms are proliferating throughout the enterprise, with each platform utilizing less than 15% of its processing capability, yet consuming 100% of its rated energy needs. The data center needs to consolidate five, ten, or more servers onto a single platform in order to enable the enterprise to grow. The IT staff needs to virtualize multiple heterogeneous applications on that server to better utilize system resources. IBM has just introduced a new member of their System p family, based on the *POWER* architecture, to do just that. The *System p 570* with *POWER6* processing may be the answer to your data center needs. To find out, please read on.

IN THIS ISSUE

- Today's Data Center 2
- IBM's POWER6 Architecture 2
- IBM's System p 570 3
- Conclusion 5

Today's Data Center

The spread of under-utilized servers is running rampant throughout the data center of every enterprise like an uncontrolled virus. If a hacker plants a virus on just one mission-critical enterprise server, and hijacks 85% of its compute cycles, the data center staff will employ a surgical team to cut that virus out and restore the server to full health. How do we explain, then, the conditions existing today where the majority of enterprise servers are operating at only 15% efficiency, while the IT staff ignores the other 85% of the CPU's capability? The cost to power these crippled servers, and cool the data center, increases the TCO of the data center and has a direct *negative* impact on the bottom-line of any enterprise. In addition, wasting natural resources limits the availability of energy to add new applications to meet the needs of a growing enterprise. *Consolidating* multiple servers onto a single platform, while *virtualizing* the environment to enable heterogeneous applications to operate, is one method of removing much of the complexity out of the IT architecture and simplifying the infrastructure to enable the staff to restore order to the data center. **While it is painful to think about redesigning the IT environment, it is irresponsible to ignore the necessity**.

Two of the most urgent imperatives any CIO must face in changing the compute paradigm of the data center are the need to change to an **energy-efficient** architecture, reducing enterprise demand on electricity, thereby improving performance per watt, and **maintaining 24x7x365 availability** to ensure continuous access to mission-critical applications. Improving the reliability, availability, and serviceability (RAS) of the application platform is not an option. Downtime is not measured in hours but in hundreds of thousands of dollars, an expense that no enterprise can afford.

Many of the servers installed throughout the enterprise were acquired in a pre-Y2K buying binge, or in the value-packed era immediately following the tech-bubble bursting in the early part of this decade. In either case, these commodity servers were designed with single-core processors from companies such as Intel and AMD, for scale-out environments running *Windows*-based applications, not for the scalability implicit in SMP architectures used in UNIX platforms and designed to grow with the enterprise. While lacking many of the RAS characteristics found in UNIX and mainframes systems, these scale-out platforms were characterized as "good-enough" for the infrastructure environments that they ran. This may be true for a single print server or file server, but when a data center installs 300 platforms in a mission-critical environment, good enough is *never* good enough. Consolidating these disparate systems requires the highest levels of consistent and reliable performance and I/O throughput, not just "good enough". They also require the flexibility to access libraries of both UNIX and Linux applications, to avoid reinventing the wheel, with the RAS characteristics necessary to ensure system availability. Finally, in order to restore simplicity to the environment, the data center needs to have a single management architecture to control the various platforms under its responsibility.

Multi-core processors have taken over the mission-critical server domain, not from Intel or AMD, but from legacy vendors who have been using UNIX servers for the past two decades, Sun with their *SPARC* architecture and IBM with *POWER*. Of these two, there is little doubt, based upon commodity benchmarks and empirical evidence, that IBM's POWER architecture has the better heritage and the better roadmap to provide energy-efficient performance.

IBM's POWER6 Architecture

POWER is not a new technology nor near its end of life. If System p were a mainframe technology, *POWER6* (*p6*) would be referred to as a mid-life kicker. Would you believe a 70-yd. Field Goal!

Dating back to 1990 with the introduction of the RS/6000 and AIX, the POWER architecture gained significant notoriety in May of 1997 when a 30-node RS/6000 with *POWER3* architecture, designed to play chess, beat World chess champion Garry Kasparov. *Deep Blue*, in 1997, had a performance rating of 11.38 GFLOPS while a single p6 microprocessor has a rating, today, of 30.5 GFLOPS. At the same time that HP is bringing their *HP-PA* architecture, with *HP-UX* operating system, to end-of-life, there is no "end game" in sight for POWER and the AIX operating system. In fact, with p6, the data center can double their performance over p5, or cut their energy con-

sumption virtually in half.

P6 is a dual-core microprocessor comprised of enough innovation to generate hundreds of new patents. It does not fall into the "me-too" trap that is prevalent in the Intel/AMD struggle for x86 supremacy. IBM has included many unique features in this next generation processor to improve performance *and* reliability, including:

- **Processor Instruction Retry** – enables hardware recovery from some non-predicted errors, by reloading a previous checkpoint and retrying in another core or another CPU;
- **Integrated Decimal Floating Point Accelerator** – a unique hardware feature included to improve decimal accuracy to comply with legal and financial rules;
- **Integrated *AltiVec Vector Technology*** – provides highly parallel operations for HPC applications with "vectorized" code;
- **On-chip Power Efficiency** – to provide for dynamic adjustment of voltage and frequency to meet processing needs and conserve energy whenever possible;
- **Ultra-High Frequencies** –provides CPU speed up to 4.7GHz per core, supporting a 7-way superscalar architecture within the same power envelope as p5+;
- **Enhanced Simultaneous Multithreading (SMT)** – to improve system performance by utilizing unused execution cycles;
- **Extended Memory Capability –** with support for up to 48GB per core;
- **Expanded Cache Capacity –** with 2x4 MB of on-chip L2 cache, and an on-chip L3 cache directory and controller; and
- **Advanced Virtualization Features –** including up to 32 virtual MAC addresses, and *Live Partition Mobility* to balance workloads across multiple servers,

The scalable performance of the p6 chip itself can best be illustrated using IBM's *rPerf* benchmark that shows the relative performance of the various POWER processors, with a POWER3 chip used as a reference point of 1.00. A 2-core p6 processor running at 4.7GHz[1] has an rPerf value of 20.13, i.e. more than 20 times the performance of the *p3*, and almost 66% faster than a 1.9GHz p5+ at 12.27. A 4-core p6 system has a value of 40.26, with an 8-core system testing at 74.89, showing the near linear scalability of the architecture. A 16-core 1.9GHz p5 HPC system has a Linpack rating of 103,100, while a 4.7GHz p6 system has a rating of 239,400, well over 2:1.

Trying to do a competitive analysis is difficult as the configurations and benchmarks vary significantly from one vendor to another. Here is one attempt. When configured as a 16-core system, the *p570* is very impressive. Using the TPC-C benchmark as a reference, the p570 has the highest ranking of any 16-core system, 1,616,162 tpmC. Furthermore, there is no competitive system even close. The #2 system in the 16-core listing is the predecessor to the p570, IBM's *p5-570*, using the *POWER5+* technology, at 1,025,170 tpmC. **That system had held the #1 position since February 2006!** To find a competitive system even close, you have to move up to the 64-core classification. HP has a 64-core *Superdome*, using a mono-core Itanium, listed at 1,231,433 tpmC. The p570 is even more impressive on a price/performance basis with a price/tpmC of $3.53 vs. $4.82 for the Superdome. Sun does not publish a TPC-C rating, so, instead, we can look at the SAP SD 2-tier ratings. Here, the 4-core and the 8-core p570s are 4 to 6 times faster.

IBM's System p 570

Between the p5's outstanding performance and energy-efficient innovations to reduce operating costs, the p570 makes an ideal system for consolidating the data center or for migrating *HP-UX* or *Solaris* applications[2] to AIX. With up to six times the memory per core of an HP *rx8640* or a Sun *v390*, and outstanding throughput, the entire p570 can actually scale with the application, not just the processor. IBM has already lowered its pricing for p5 and p5+ systems to compete more aggressively with HP's *rx7640* Itanium servers, and they are pricing the p570 to compete with HP's *rx8640 Itanium* servers.

The p570 is highly scalable. The IT staff can even add nodes to the configuration without

[1] The p6 processor is also available at 4.2GHz and 3.5GHz, and can scale to 128 threads.

[2] Migrations are currently a major effort at IBM with over 430 UNIX migrations already complete, over 80% of them from HP or Sun platforms.

taking this highly reliable system down. This includes nodes for expansion or a node that is being returned to active status after maintenance. Reliability is another factor that can have a positive, or negative, impact on TCO. Downtime is costly! The p570's design takes advantage of IBM mainframe inspired RAS innovation to enable continuous availability, helping to eliminate planned downtime and reduce any unplanned outages. A reliable system will enable the enterprise to maintain market competitiveness, helping to ensure customer loyalty.

In order to reduce the TCO of the data center and to conserve our environment, IBM has implemented an extensive set of energy saving innovations within the p570 EnergyScale architecture, enabled through IBM's *Power-Executive*, include:

- Digital Thermal Sensors;
- Variable fan speed;
- Rear door heat exchanger;
- Optional Advanced POWER Virtualization – enabling Micro-Partitioning, Partition Mobility, and VIOS V1.4 to share physical I/O resources between partitions, among others; and
- Utility CoD to provide, autonomically, temporary processor capacity within the shared processor pool (3Q07), granular tracking, and usage-level billing.

PG&E has recognized the value of IBM's Energy-Scale technology, offering energy rebates to customers wishing to take advantage of the p570's high-energy efficiency.

IBM has already announced an upgrade from the p5-570, and there are planned upgrades from both the p5-590 and p5-595.

Hardware Configurability

Each 4U p570 node can support from 1-4 dual-core p6 processors, currently available at 3.5GHz, 4.2GHz, and 4.7GHz clock speeds. With two threads/ core, each node can support 16 threads. A fully configured four-node p570 system can support up to 16 sockets and 64 threads. This will assure any mid-sized to enterprise data center of having sufficient processing power to consolidate their mission-critical applications.

With each core in a p6 CPU having access to up to 48GB of high performance *Chipkill* DIMM memory; a single p570 can support 192GB, depending on memory speed. A four-node configuration can support up to 768GB. This will ensure that with up to 10 partitions per core available, each partition in a virtual p570 environment has enough memory to scale with the application[3]. In addition to the RAM, each p6 core has 4MB of L2 cache, with an additional 32 MB of L3 cache shared between the cores.

Consistent with the scalability of processing and memory, I/O is equally scalable with over 300GB/s of processor I/O bandwidth per CPU and six PCI slots per node, four PCI Express and two PCI-X, for high performance I/O. In addition, there are up to eight optional I/O drawers per node, with 32 per system. Each p570 node comes with SAS disks (up to six bays/node), supporting up to 1.8TB of disk internally. Optional external storage scales that to over 30TB. Existing HMCs[4] can manage both POWER5 and POWER6 systems, enabling the data center to simplify the infrastructure.

Software (AIX/Linux)

The p570 is available with *AIX 5L* V5.2 or 5.3 and *SUSE Linux Enterprise Server* 10 SP1. *Red Hat Enterprise Linux 4.5 for Power* will be available in 3Q07. *AIX Version 6.1* will be available in 4Q07 and is binary compatible with existing applications[5]. The AIX POWER Hypervisor enables the p570 to run a heterogeneous environment, with AIX and Linux applications sharing processing, memory, and I/O resources.

For the first time, IBM will make an early beta version of *AIX 6* available during 3Q07, probably in July. This will give their clients an opportunity to gain valuable early experience with this new version of IBM's open, standards-based UNIX O/S in data mining, database processing, transaction processing, and high performance computing environments. The new release will include significant new capabilities for virtualization, security, continuous availability, reliability, and management.

One of the new ***virtualization*** innovations is

[3] This gives *p570* users a significant advantage over x86 products such as VMware that limit a partition to 4GB.

[4] Hardware Management Consoles.

[5] Applications will perform better if they are recompiled to take advantage of p6 hardware innovations such as Decimal Floating Point.

Workload Partitions (WPAR), an architecture that complements the existing IBM logical partitioning (LPAR) by reducing the number of AIX images that have to be managed when consolidating workloads. WPAR enables the IT staff to consolidate multiple applications within a single instance of AIX 6. In addition, *Live Application Mobility* enables AIX to move a WPAR from one system to another without restarting the application. This facilitates the capability of the data center to control planned downtime to do upgrades or maintenance on a server without disrupting the end user, or shutting a system down completely to conserve energy during periods of light activity. Live Application Mobility will be available in IBM's *Workload Partitions Manager*, scheduled to be available as an option to AIX 6, with simplified installation and configuration features to manage WPARS across multiple systems.

AIX 6 will have several significant enhancements to ***secure*** the System p environment, including:

- *Role based security* – to enable administrators to grant authorization for management of specific AIX resources to users other than root according to their function;
- *B1 Trusted AIX* – implemented as an option to meet critical government and private industry security requirements; and
- *Filesystem encryption* – enables the encryption of IBM Journaled Filesystem Extended (JFS2) data to prevent the compromise of data even to root level users.

In order to improve the ***reliability*** of the AIX operating system, IBM has enhanced AIX 6 with several features to ensure ***continuous availability***. These include:

- Kernel support for *Storage Keys* – to protect key portions of the AIX O/S and application s/w from accidental memory overlay that do not require a system reboot, reducing unplanned outages;
- *Concurrent kernel update* – to deliver some kernel updates as interim fixes; and
- *Dynamic tracing* – simplifies the debugging process. It allows the user to insert breakpoints in existing code without having to recompile.

Among the new ***management*** features that will appear with AIX 6 is the *Power Executive*. Power Executive will control:

- Thermal/Power Measurement;
- System Health Monitoring/Maintenance;
- Power Capping; and
- Power Saving.

Conclusion

It should come as no surprise that yesterday's low-cost, open systems servers use yesterday's technology. They are wasting energy, today. Do you want to continue wasting energy tomorrow?

Without a doubt, IBM's POWER6 microprocessor is the fastest POWER processor yet. It is also simply the fastest commodity chip available. Combined with IBM's innovation in energy-efficiency, virtualization, continuous availability, and security, the p570 provides an ideal platform to consolidate your data center and migrate away from yesterday's solutions. If you are looking for a scalable server to solve your business growth problems, look no further. With thousands of AIX and Linux applications at your disposal, the p570 can improve your performance and lower your TCO. Not only can the p570 help improve profitability to the bottom line; it can also help to slow the electric meter that governs your data center.

With a solid history and a healthy roadmap, IBM's POWER architecture and System p server family may be the answer to your enterprise's problems as well.

About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

David Reine **is Director, Enterprise Systems for The Clipper Group.** Mr. Reine specializes in enterprise servers, storage, and software, strategic business solutions, and trends in open systems architectures. He joined The Clipper Group after three decades in server and storage product marketing and program management for Groupe Bull, Zenith Data Systems, and Honeywell Information Systems. Mr. Reine earned a Bachelor of Arts degree from Tufts University, and an MBA from Northeastern University.

- ***Reach David Reine via e-mail at dave.reine@clipper.com or at 781-235-0085 Ext. 123. (Please dial "123" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group Captain's Log**, **The Clipper Group Voyager**, **Clipper Notes**, and *"clipper.com"* are trademarks of The Clipper Group, Inc., and the clipper ship drawings, *"Navigating Information Technology Horizons"*, and *"teraproductivity"* are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.