



# **Exploring @server pSeries 650 and pSeries 660-6M1 Performance Attributes**

**Bret Olszewski**  
**[breto@us.ibm.com](mailto:breto@us.ibm.com)**

**November 12th, 2002**

## Introduction

With the debut of the IBM @server™ pSeries™ 650, IBM recaptures the mantle of technology and performance leadership in the eight-way server arena. Benchmark results have been submitted for approval to the Standard Performance Evaluation Council (www.spec.org) in November 2002. These prove eight-way leadership, including SPECweb99, SPECjbb2000, SPECsfs97, SPECfp\_rate2000 and SPECCompM2001 (for details, refer to Table 3). The p650 constitutes a very different offering than predecessor products, such as the pSeries 660 Model 6M1. This paper explores some of the performance differentiators between the p650 and the p660-6M1.

## The p660-6M1

The pSeries p660-6M1 similarly vanquished the competition when it stormed onto the scene in October of 2001. It set high watermark performance results on TPC-C, SPECweb99, and SPECjbb2000. This impressive set of results set high expectations for successor products.

With the excitement associated with the introduction of the POWER4™ processor-based pSeries 690, little attention was focused on describing and exploiting the interesting technology associated with the p660 Model 6M1. Rather, this machine was known primarily on the basis of its raw performance and scalability. But, in order to understand the differentiators of POWER4 technology, it is necessary to revisit the p660-6M1.

The microprocessor in the p660-6M1 is usually referred to by its external designation, RS64 IV. This processor is actually the fourth technology iteration of one that first appeared in 1998 in the RS/6000® Model S7A [BOR2000]. Within IBM, we refer to these microprocessors as the STAR family, as all of the technology iteration code names included the word STAR within them. Over time, technology enhancements occurred to the basic microprocessor which allowed the frequency and off-chip caches to increase at a rapid rate as shown in Table 1. This processor, in its many variations, powered a number of very successful pSeries and IBM @server iSeries™ models.

In a sense, this microprocessor's design was influenced by earlier RISC designs such as POWER and POWER2™. It included what could be simplistically viewed as a 4-stage pipeline with an in-order execution

design that was capable of dispatching up to four instructions per cycle. Instructions are fetched, decoded, dispatched, and completed. At dispatch, instructions are put into "pipes" for execution. These pipes are the upper limit on the number of instructions that the microprocessor can perform at any one time. The four pipes and their function are described in Table 2.

**Table 1 - STAR family in RS/6000 & pSeries systems**

Year	Frequency	L2 Cache	Model
1998	262 MHz	4MB	S7A
1999	340 MHz	4MB	H70
1999	450 MHz	8MB	S80
2000	500 MHz	4MB	M80
2000	600 MHz	16MB	p680
2001	750 MHz	8MB	p660-6M1

**Table 2 - STAR execution pipes**

Execution unit	Function
B-pipe	Branch instructions
M-pipe	Simple and complex integer instructions
R-pipe	Simple integer instructions
S-pipe	Load/store instructions

The M-pipe actually does double duty. In addition to processing integer operations, floating-point computation instructions are also dispatched there. In most commercial workloads, the dispatch to units is disproportionate with the S-pipe typically seeing approximately 40% of all instructions dispatched to it. Because of this and the simple in-order dispatch, it is fairly rare that four instructions are dispatched in any cycle, though two and three instruction dispatch is common.

As a microprocessor executes instructions, it is frequently challenged with conditional branches. A conditional branch causes the flow of instructions to go down one of two paths. Since the processor is most efficient if it doesn't wait to find out exactly which path will be taken, most guess at the right path. If they are wrong, partially completed instructions may be thrown away. But with reasonably good guessing, performance is improved. The guessing is most often based on keeping track of what conditional branches have recently been executed and which direction they typically take. These history-based schemes are generally referred to as dynamic branch prediction. Unlike many of its contemporaries, STAR had no dynamic branch prediction. Rather at each conditional branch, the processor assumed that the branch was not

taken and instructions were dispatched accordingly. Instructions for the non-guessed path were fetched as well. When the branch was resolved that was incorrectly speculated, it was fairly inexpensive to discard the speculated instructions and begin execution down the correct path. This simple approach was fairly effective in the design, since it had a fairly short pipeline. This, coupled with a robust instruction cache design did not tend to make the lack of dynamic branch prediction a limiting factor for codes in its target markets. But it did have the ramification that software performed best when tuned such that conditional branches were not taken more often than they were taken.

Since the processor was not intended for scientific and technical computing, it was not endowed with extraordinary floating-point performance. A single floating-point unit is employed. Unlike POWER3™ [FOC2000], no hardware prefetch mechanism is implemented. This, coupled with the limitation of a single load/store at a time execution through the S-pipe, limited the memory bandwidth that could be achieved.

A number of STAR-based pSeries systems were enabled to exploit hardware multi-threading (HMT). HMT is a feature of STAR that provided a mechanism for hiding memory latency by having each microprocessor support execution of two “threads” of instructions. When one was blocked by a long latency memory operation, the other could continue to execute. The benefits of HMT have been well documented in previous publications [OLS2002].

The implementation of the caches was state-of-the-art IBM technology. In the p660-6M1, the on-chip level 1 (L1) instruction cache was 128KB in size. The on-chip L1 data cache was also 128KB in size. The directory for the level 2 (L2) was also on-chip, allowing the latency for an L2 cache hit to be as low as nine processor cycles.

While the STAR microprocessors were used with many different memory controllers, the p660-6M1 gave it the benefit of recycling the high-end technology from the RS/6000 Model S80 and p680 into the mid-range. Essentially, the p660-6M1 benefited from a crossbar switch-based memory controller which actually provided more bandwidth than could normally be used by STAR processors. I/O bandwidth was also ample when the system was configured with sufficient I/O drawers.

While not embodying many of the elements of other microprocessors such as HP’s PA-8000 [KUM1996], the STAR family produced solid performance. In fact, it was designed to run TPC-C-like workloads and succeeded beyond the expectations of its original designers. The secret of this performance was primarily tied to the large, fast on-chip caches and some leading edge off-chip L2 cache technology. The effect of these caches was to reduce the average time that the processor was stalled waiting on memory access.

## **POWER4 - The next generation**

In the late 1990’s, IBM had one microprocessor family to address the scientific and technical market with POWER3 and another to address the traditional commercial market with STAR. While the STAR family served its purpose admirably, it was apparent that its design point couldn’t be continued indefinitely. The convergence of traditional commercial and scientific and technical computing provided an imperative to consider a single microprocessor which could address both markets. The requirements of a processor for the 21st century included the ability to scale to higher frequencies, higher memory bandwidth, and greatly increased floating-point rates dictated a more radical microprocessor design. To achieve these goals, many of the basic tenants of the STAR design could not be continued. So the decision was made to implement POWER4. This microprocessor inherits some of the characteristics of POWER3, but with a completely new design.

Since the design requirement anticipated scaling to high frequency, a much more deeply pipelined approach was required. Instructions are pipelined between sixteen and twenty-one stages. This pipelining allows many instructions to be in some phase of execution concurrently within the core. In fact, up to 100 instructions could be in flight at the same time. Deep pipelining means that branch prediction is critical, since canceling instructions is more expensive in CPU cycles of latency than on STAR processors. Prediction occurs for both branch direction and branch address. Multiple branches can be predicted each cycle with provisions for up to sixteen predicted branches in concurrent execution. Speculative execution of instructions occurs with out of order execution. The speculative efficiency of the processor is highly dependent on the effectiveness of branch prediction.

The POWER4 chip was the first general purpose implementation to include two processors on the same

die. This design optimizes packaging for large symmetric multiprocessor (SMP) systems in a number of ways. First, it reduces the physical area required for a fixed number of microprocessors. Physical size is important in maintaining the speed of the buses which connect the processors to each other and memory. Second, it reduces the electrical “load” on the buses, making it easier to run them at very high speed. Third, it allows very efficient sharing of information between the two processors on the chip via cache.

Starting from a high frequency design point, a very different memory hierarchy design is required when compared with the STAR series systems. Since high speed access to the L1 cache is required, it was not an option to merely have large L1 caches on the chip. Rather, the L1 caches are made as large as possible without affecting cycle time. It is very important that the load latency for cache hits remains one cycle. The L1 caches are supplemented by an on-chip L2 cache. The L2 cache has higher latency than the L1 cache, but it is significantly faster than using an off-chip L2 cache, since the speed and access times of off-chip caches is falling behind the performance of microprocessors. Other designs, such as the Intel® Itanium® 2 [KRE2002] have actually gone to as many as three levels of on-chip cache. Because the L2 cache is on the microprocessor chip, the amount of cache is limited by the die size. While the L2 cache is faster in nanoseconds than the L2 cache used in STAR, it is smaller. To compensate for the L2 cache size, all pSeries POWER4 processor-based systems include a third level (L3) of cache which is off-chip. The L3 cache has relatively longer latency than the L2 cache.

POWER4 systems are capable of driving much higher memory bandwidth than STAR systems. To this end, the memory subsystem runs at more than three times the clock speed of the p660-6M1. The memory controller is capable of handling more simultaneous requests, driven by the voracious requirements of the POWER4 prefetch mechanism. But the latency, or time measured in nanoseconds to access a piece of data in memory has not improved significantly over the p660-6M1.

The p650 is the first system to contain the second generation POWER4 processor known as POWER4+™. The POWER4+ microprocessor improves on its predecessor by virtue of a technology improvement that allows it to run at higher frequency with less power consumption and dissipation. A number of other minor improvements, including a slightly larger on-chip L2

cache and revised memory queuing algorithm, also generally improve system performance.

## Where the rubber meets the road

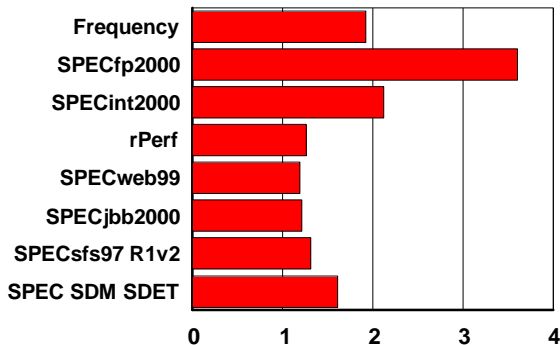
Architecture is fascinating, but the true value of microprocessor technology is shown in how well it runs programs. The UNIX® world has long depended on industry-standard benchmarks to compare and contrast performance between architectures and systems. Table 3 shows p650 benchmark results submitted to SPEC in November 2002.

**Table 3 - p650 1.45 GHz 8-way benchmarks**

Benchmark	Result
SPECweb99	12,400
SPECjbb2000	114,892
SPECfp_rate2000	82.4
SPECCompM2001 peak	9,694
SPECsfs97 R1.v3 UDP	55,825
SPECsfs97 R1.v3 TCP	55,526
SPECsfs97 R1.v2 UDP	71,075
SPECsfs97 R1.v2 TCP	70,894

These benchmarks are useful indicators of performance, but caution must be exercised before depending on them. When comparing benchmark results, it is important to know differences in software levels and tuning that can result in benchmark results which are not a pure indicator of hardware performance differences. Chart 1 shows some AIX 5L™ workload performance ratios between p650 and p660-6M1 on a number of workloads. These ratios have reasonably pure software and tuning conditions, so these represent primarily hardware performance. The rPerf metric is a synthetic metric provided by IBM for customer use in comparing the relative commercial performance of pSeries systems.

**Chart 1 - p660-6M1 to p650 speedup**  
 Ratio of speed: p650 (1.45 GHz) to p660-6M1 (750 MHz)



The first thing that demands attention in the chart is that the relative gain is not easy to put into a single number. Rather, the benefit of p650 depends tremendously on the nature of the workload. But with a little bit of analysis, the data begins to make sense.

First, the improvement in SPECfp2000 is fairly easy to internalize. As described previously, the STAR family of processors were not designed for intensive floating-point applications. Most of the codes in this workload are not particularly memory intensive, so the ability to execute more instructions concurrently, coupled with higher frequency results in better than frequency performance scaling.

Second, the improvements in SPECint2000 are primarily in frequency. Additionally, its more sophisticated core shows the ability to exploit more instruction-level parallelism than the simpler STAR microprocessor. Since most of the programs in this benchmark suite are not particularly memory intensive, the POWER4+ is able to scale above frequency.

Third, workloads such as SPECweb99, SPECjbb2000, and SPECsfs97 have bigger memory footprints and can emphasize the memory subsystem more than the microprocessor core. In fact, SPECweb99 and SPECsfs97 are very similar in that they are using vast quantities of memory for caching objects. The access to these objects is semi-random, which results in rather poor memory locality. Because the cache efficiency is limited by the randomness of the data access, system performance tends to be gated by memory access performance. Comparing workloads can also introduce the fact the p660-6M1's support of HMT allows it to hide a considerable amount of memory latency in many environments. The SPECweb99 results on p660-6M1

include the use of HMT, while the SPECsfs97 and SPECjbb2000 do not.

Finally, the SPEC SDM SDET workload is a good example of an environment which mostly fits in the L3 caches of the p650 system. This now obsolete SPEC workload is composed of a set of scripts which are used to drive operating system workload. Because the commands executed run a very short time on today's systems, the programs tend to be created, run, and terminate within the caches of the system. Essentially, they hardly "spill" out of cache to memory. Since the workload is less memory intensive than SPECweb99, performance scales better with frequency.

So why can't the highly pipelined, speculatively dispatched POWER4+ system overcome memory latency? The answer is that it can do so, but only to a point. Consider that typical load/store instructions are pipelined eighteen stages deep. Since a data access which is resolved in the L2 cache is on the order of twelve cycles, the pipeline can more often than not "hide" some or all of the latency of an L2 cache hit. But, the latency to memory is many times more than of the L2 cache, so the processor will tend to stall on accesses that go off-chip. But, if accesses to data are sequential, the microprocessor will begin to prefetch the data from memory and hide the memory latency. Unfortunately, most commercial workloads do not have a high degree of sequentiality to their memory accesses. Because of this, memory latency is usually the limiting factor on microprocessor performance.

### But wait, there's more

Benchmarks are generally useful indicators of performance, but there are other attributes of hardware and software that are undervalued in many of today's commercial benchmarks. For example, in the commercial workload world, memory bandwidth is not usually considered to be critical. Indeed, most applications such as online transaction processing (OLTP) don't explicitly require enormous memory bandwidth, many everyday tasks can benefit from higher bandwidth. Consider the simple act of reading a large file using a filesystem. Under normal circumstances, the file data is cached in system memory. That means that moving data from the data to the user application involves copying. So, a system which has higher memory bandwidth and memory prefetching will take less CPU time, even if the elapsed time for the operation is limited by disk performance. When this is taken into account, the fact that p650 can

deliver more than five times the memory bandwidth to a processor than the p660-6M1 can, opportunities exist to find ways to use it. Memory bandwidth is also useful when moving data from one cache to another.

A careful evaluation of many popular benchmarks, such as TPC-C, show that most access data with an even distribution. That is, the probability that a transaction will access one record in a database table is more or less the same as accessing any other record. Normally, even distribution of data is desirable in benchmarks. It helps to make the benchmark stable. Multiple runs of the workload will result in similar, predictable performance. It also makes it more difficult for vendors to “optimize” the workload in ways that defeat its original intentions. But even distribution of data is not the norm in the real world. Consider an airline booking system. Is the number of flights departing daily from Duluth, Minnesota the same as Chicago, Illinois? No, the probability of accessing data associated with Chicago is higher, and not just by a little. In a running airline system, one could easily envision records associated with Chicago will be bouncing back and forth from processor to processor rapidly, while Duluth is infrequently accessed. The shared L2 cache design of POWER4+ can frequently assist in this task, as data used by one processor on the chip can be supplied to the other processor very rapidly. The p650 memory subsystems can also speed hot data between the processors efficiently.

Another arena that the p650 shows benefit is in I/O bandwidth. Comparing with the p660-6M1, both fully configured with I/O drawers, the p650 is capable of up to twice the available I/O bandwidth. This can result in significant benefit, particularly for systems with high-speed SAN and/or Gigabit Ethernet adapters that can tax I/O subsystems.

## Conclusions

The performance of the p650 provides leadership for pSeries. The improvements in performance of the p650 over the predecessor product, the p660-6M1, are significant in numerous dimensions. The fact that performance does not scale at the rate of frequency for most workloads is normal and is expected behavior for commercial systems.

## References

- [BOR2000] J. Borkenhagen, R. Eickemeyer, and R. Kalla. A Multithreaded PowerPC Processor for Commercial Servers, IBM Journal of Research and Development, November 2000, Vol. 44, No. 6, pages 885-898.
- [OLS2002] Bret Olszewski, Octavian F. Herescu. Performance Workloads in a Hardware Multi Threading Environment. Fifth Workshop on Computer Architecture Evaluation using Commercial Workloads, February 2 2002, pages 68-79.
- [KUM1996] Ashok Kumar. The HP PA-8000 RISC CPU. Hot Chips VIII, August 18-20 1996, pages 9-20.
- [FOC2000] F. P. O’Connell, S.W. White, POWER3: The next generation of PowerPC processors. IBM Journal of Research and Development Vol. 44 No. 6 November 2000.
- [KRE2002] Kevin Krewell. Itanium 2 Arrives With a Benchmark Bang. Microprocessor Report, August 2002.

Copyright IBM Corporation 2002

IBM Corporation  
Marketing Communications  
Server Group  
Route 100  
Somers, New York 10589

Published in the United States of America  
11-02  
All Rights Reserved

This publication was developed for products and/or services offered in the United States. IBM may not offer the products, features, or services discussed in this publication in other countries. The information may be subject to change without notice. Consult your local IBM business contact for information on the products, features and services available in your area.

All statements regarding IBM's future directions and intent are subject to change or withdrawal without notice and represent goals and objectives only.

IBM, the IBM logo, the e-business logo, @server, AIX 5L, iSeries, POWER2, POWER3, POWER4, POWER4+, pSeries and RS/6000 are trademarks or registered trademarks of International Business Machines Corporation in the United States or other countries or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Intel and Itanium are registered trademarks of the Intel Corporation in the United States or other countries.

Other company, product, and service names may be trademarks or service marks of others.

IBM hardware products are manufactured from new parts, or new and used parts. Regardless, our warranty terms apply.

Photographs show engineering and design models. Changes may be incorporated in production models.

Copying or downloading the images contained in this document is expressly prohibited without the written consent of IBM.

This equipment is subject to FCC rules. It will comply with the appropriate FCC rules before final delivery to the buyer.

Information concerning non-IBM products was obtained from the suppliers of these products. Questions on the capabilities of the non-IBM products should be addressed with the suppliers.

All performance estimates are provided "AS IS" and no warranties or guarantees are expressed or implied by IBM. Buyers should consult other sources of information, including system benchmarks, to evaluate the performance of a system they are considering buying.

The IBM home page on the Internet can be found at [www.ibm.com](http://www.ibm.com)

The pSeries home page on the Internet can be found at [www.ibm.com/servers/eserver/pseries](http://www.ibm.com/servers/eserver/pseries)