



Cell Processor Based Blade Server White Paper

May 13th, 2005

The Cell Processor

Background:

Modern processor architectures have many similarities. A central processing unit (CPU) interprets and executes instructions. A memory subsystem stores programs and data. An Input/Output (I/O) subsystem handles the movement of data and programs between the CPU and the external world, be it local or networked. Special functional units to provide additional functions, like security or special graphics output, are sometimes added to the CPU or the I/O subsystem. To improve memory bandwidth, memory subsystems become more complex with multiple levels of cache.

Processor speeds improve over time, as technological improvements in manufacturing allow more transistors per unit of chip real estate, closer packing of components, etc. These improvements generate more heat and consume more power, so strategies must be implemented to address these increases.

Moore's law says that we can expect to double the transistor density every 18 to 24 months. Historically, that has been the case pretty much since the commodization of the integrated circuit.

The challenge is to keep a balanced system. Memory speed (i.e. latency and bandwidth) improvements do not keep up with processor speed, so we introduce larger, multiple level caches with exotic management strategies. Pipelines become deeper. Some of those new transistors that Moore promised are used to address these issues, and processor efficiency improves slower than transistor density.

Improved manufacturing techniques, like 7 Level Copper Silicon On Insulator (SOI) or Low K (impedance) dielectrics help, but the inefficiencies remain. As components like gate dielectrics approach real physical limits (i.e. a few layers of atoms), a radical improvement in microprocessor design is needed.

Cell Processor History:

The Cell architecture is an outgrowth of a strategic alliance between Sony Computer Entertainment Inc®, Toshiba Corporation®, and IBM Corporation®, formed in 2001. The intent was to develop a Cell processor, based on IBM's Power Architecture™, that would provide very significantly better performance in the games, media and digital content market. The STI (Sony, Toshiba, IBM) design center was opened in 2001 in Austin, and development began. The first technical disclosures were in February 2005 at the International Solid-State Circuits Conference.

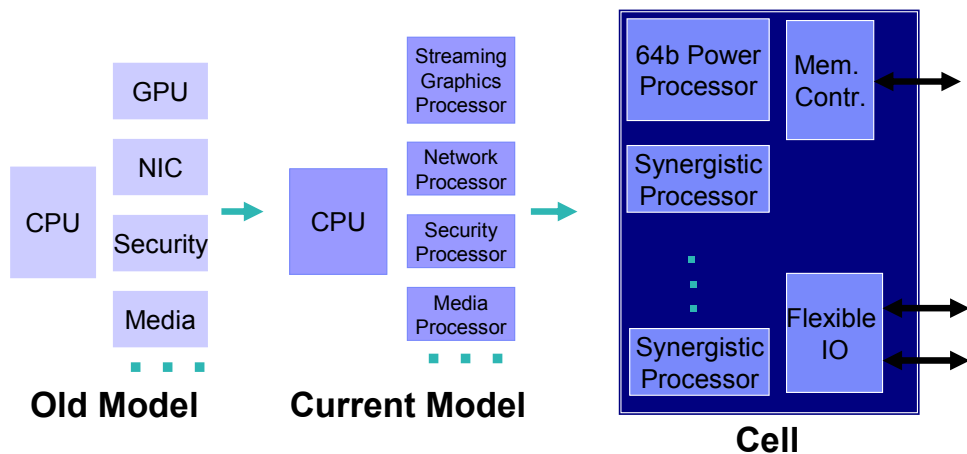
Cell Processor Architecture:

While conserving the base architecture of traditional systems, the design and implementation of the Cell architecture addresses the pitfalls of traditional microprocessor architecture, that of simply trying to provide more of the same just at higher clock speeds:

Next generation processors address programming complexity and trend towards programmable offload engines with a simpler system alternative:



Next Generation Processors address Programming Complexity and Trend Towards Programmable Offload Engines with a Simpler System Alternative



The Old Model represents microprocessor designs of previous generations, the Current Model represents contemporary design. Note the improvement from the old model to the current model. Functions that were adapters have become processors. The CPU is offloaded, natural parallelism is achieved, and functionality is added with embedded processors on the adapters.

The Cell processor is a multicore design, including a 64 bit processor element called a PPE core, 8 synergistic processors called SPEs, a memory controller, and I/O control in the same package.

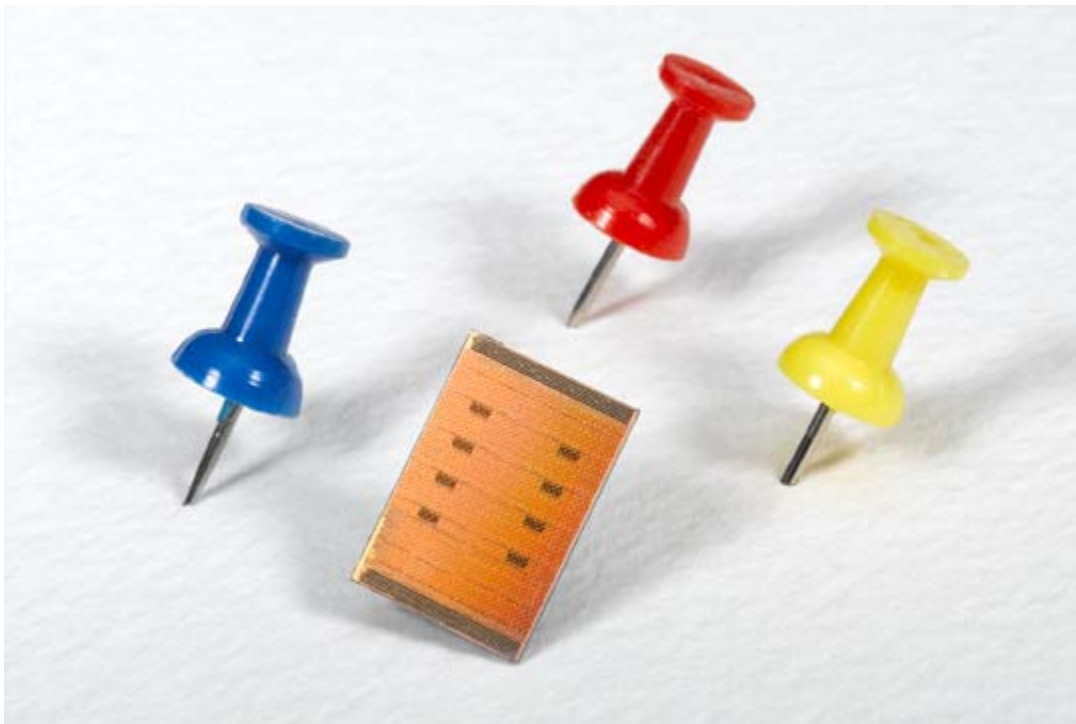
The PPE is a 64-bit Power Architecture which, along with the 8 SPEs is a 10 way coherent multiprocessor with respect to access to main memory. The chip has 2.5MB of on- chip memory (512KB L2 and 8 * 256KB on the SPEs). The main memory is a

Rambus™ design, and may be up to 1GB XDRAM depending on the memory chip configuration chosen.

Each of the Synergistic Processor Elements (SPEs) has a 128 entry 128 bit register file with 256KB local memory. Each operates as a Single Instruction Multiple Data (SIMD) accelerator with up to 16 way SIMD capability. SPEs are RISC architecture, especially designed for the Cell architecture, and allow 16 concurrent memory accesses (to the local store) per SPE via DMA.

Both the on-die memory and I/O controllers are Rambus design.

In prototype, the initial Cell has 234 million transistors and a die size of 221 square millimeters. Physically, it looks like this:



This combination of a fairly standard RISC processor with 8 attached SIMD processors is quite unique, and will allow certain classes of operations to be executed several times faster than the PPE alone.

The Cell Processor Based Blade Server

Hardware:

Cell processors will be aggregated in an IBM BladeCenter™ package for high power dense servers. When packaged as a blade, the Cell Processor Based Blade will be a double width blade, allowing 7 Cell processor based blades to fit into a standard IBM BladeCenter chassis. The Blade will have a 2 way SMP Cell processor.

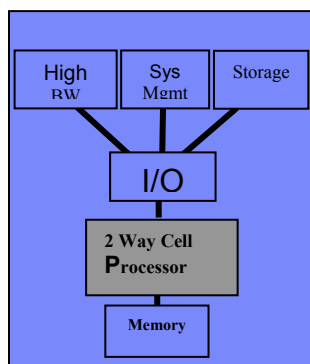
Each of the Cell Processors will have Support Logic as follows:

- 1 Power Architecture™ based core, or PPE
- 8 Synergistic Processor Elements (SPEs)
- Single SMP OS image

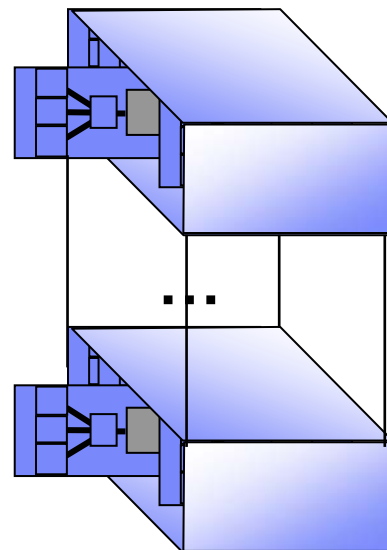
The Blade will have:

- 1GB XDRAM, with current memory packaging
- 2 optional PCI-Express attached InfiniBand™ cards

The BladeCenter Chassis will have updated management module firmware, and provide external InfiniBand switches with optional fibre channel (FC) ports. The BladeCenter chassis also provides power, and Sense Logic Control firmware to connect processor & support logic to the service processor.



**Cell Processor
Based Blade**



**16 TFlop rack (at 3 GHz)
(6 BladeCenter Chassis)**

Software:

The Cell Processor Based Blade Server will run Linux™ with extensions to utilize the data parallel SPEs. Each blade will have one OS image. Each blade is a 2-way SMP. Given that the base software is Linux, we can expect that middleware and applications that run in the Linux environment will run properly here. Applications compiled for the Power Architecture or the PowerPC® ISA are binary compatible with the PPE. Initially, compiler extensions will allow programmers access to intrinsics that utilize the SPEs. Over time more automatic parallelization may be provided.

With basic building blocks of "doublewide" blades and the IBM BladeCenter Chassis, configurations are pretty straightforward. One to seven Cell Processor Based Blades may be installed in each BladeCenter chassis, allowing from 2 to 14 Cell processors. Since a BladeCenter chassis is 7U high, 6 BC chassis will completely fill a standard 42U rack. Communication between blades within a chassis and between blades in different chassis will be by Infiniband or Ethernet. Communications with the outside world will be the same. Here a Blade is shown in prototype:



Performance:

Performance will be released at a later date. Preliminary indications are that a single SPE can do the work of a standard Power Architecture based microprocessor in cases where the algorithms are suitable. A cell, in this case, might have the power of 10 traditional processors.

Current projections show peak single precision floating point of 64GFlops per GHz. Application performance will be released later. Preliminary tests on the prototypes indicate a high degree of efficiency with regard to realized performance compared to theoretical peak performance for target applications. The actual operating clock speed will be system and power budget dependent.

Algorithms that lend themselves to data parallel structures will show the most speedup. Since the compilers supplied are enhanced to provide access to the SPEs where appropriate, recompilation and code restructuring are necessary to see any improvement in execution time.

The architecture of the Cell Processor Based Blade Server is inherently scalable. Blades within a BladeCenter chassis communicate through the chassis infrastructure. Since only 7 Cell Processor Based Blades may be put into a BladeCenter chassis, internal latencies are quite low. Addition of multiple BladeCenter chassis in a rack adds switching components, allowing uniform latencies and bandwidths to be maintained to arbitrary large configurations.

Applications:

The Cell architecture was initially conceived for acceleration of digital content creation. Because this is an extremely computationally intense problem, engines optimized for this have utility in many other areas.

First, let's distinguish between classes of graphics used in games and movies. Game systems use polygon based, real time, lower quality graphics. While Cell technology might be used for the geometry engine, the rasterization will likely be done by dedicated graphics chips.

The second type of graphics is photo realistic, done offline. Studios today use large clusters of rack dense servers, including blade servers, for this purpose. These images are created entirely in software and may take minutes to hours per frame. Cell Processor Based Blade Servers can improve the performance of this compute intensive task by an order of magnitude.

Games represent the holy grail of computational algorithms in a sense, and techniques used in games often have applicability in other scientific areas. A classic pathfinding algorithm, called A* (pronounced A-star) is used extensively in games and military

simulation. It is very CPU intense, and becomes more so exponentially with the size of the terrain. Cell processors are particularly effective at A* because of the SPE's register structure and parallel execution. Interestingly enough, monte carlo hedge fund simulation is a similar problem and should benefit from the Cell architecture as well.

SPEs have instructions specifically for scatter/gather operations. Coupled with the ability to have multiple I/O operations outstanding, any algorithms heavily dependent on scatter/gather should demonstrate significant speedup. Life sciences applications like BLAST may expect significant speedups from this architecture.

The very robust communication structure between SPEs is optimal for algorithms where computation is performed on a node, the results passed to a nearest neighbor, more computation is performed, etc. Any application that uses Fast Fourier Transforms successively is a natural for the Cell architecture. Signal processing (like seismic) and computational fluid dynamics applications should benefit significantly from the Cell.

Less traditional applications, like video surveillance, are a perfect fit for Cell processors. The video encoding capabilities, coupled with an Artificial Intelligence engine, might recognize problems and generate appropriate alerts.

The applications list above is not exhaustive, but is meant to give a feel for the breadth of applications that will perform better on Cell architecture systems.

Summary

The Cell processor ushers in a new era of capabilities for Digital Media, and by extension many other High Performance Computing applications. It is a multi-core multi thread design that promises to be faster than the Intel Architecture (IA) and the various RISC solutions available in the market today. The core CPU is based on IBM's Power Architecture, in this design called a Power Processor Element, and it also includes 8 Synergistic Processor Elements for each PPE, operating in a data parallel mode.

Given the familiarity of the base environment, Linux on Power, the Cell processor will feel familiar to most users. Recompile and perhaps restructuring will be needed to take advantage of the SIMD SPEs, and the resultant bandwidth and computational power may range to as much as an order of magnitude greater than a single Power CPU.

Packaging of the Cell Processor Based Blade Server is the familiar BladeCenter chassis, with changes in the firmware to manage the Cell processor components. Each blade will be a double-wide, with two Cell processors and associated support components. A single die will house the PPE, the 8 SPEs, a memory controller, and an I/O interface built on the same die.

Assuming an estimated clock speed of 3 GHz, a fully populated BladeCenter represents 2.7 Tera-Flops of theoretical power for the SPEs alone. A fully populated rack could be 16.5 Tera-Flops. Actual benchmark data will be available at the formal announcement of the product. Production clock speeds will be gated by system and power constraints.

With Cell processors being as much as 10 times the power of traditional processors, we have the ability to bring very high computational resources to bear with much lower power and heat. The BladeCenter packaging furthers the goal of very high processing power with very low environmentalals.

Acknowledgements:

I was fortunate to have access to many competent and willing people, especially:

Dinh Pham, of IBM's Deep Computing organization for project management and encouragement.

Numerous technical resources like:

James A Kahle, IBM Fellow
Ted Maeurer, STI Design Center
Barry Minor, STI Design Center
Peter Hofstee, STI Design Center
Joyce Fitz Ruff, IBM STI Design Center
Roland Seiffert, Boeblingen Development Laboratory
Toshi Sanuki, Systems Development Laboratory, IBM-Japan
Bruce D'Amora, IBM Research
Ashwini Nanda, IBM Research
George Dolbier, Sr. IT Architect, SMB

Authored By:

NormSnyder LLC
Norm Snyder, President
NormSnyder@ieee.org



© IBM Corporation 2005
IBM Corporation
Systems and Technology Group
Route 100
Somers, New York 10589

Produced in the United States of America
May 2005
All Rights Reserved

This document was developed for products and/or services offered in the United States. IBM may not offer the products, features, or services discussed in this document 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 future directions and intent are subject to change or withdrawal without notice and represent goals and objectives only.

IBM, the IBM logo, POWER, Power Architecture, PowerPC, BladeCenter are trademarks or registered trademarks of International Business Machines Corporation in the United States or other countries or both. A full list of U.S. trademarks owned by IBM may be found at:
<http://www.ibm.com/legal/copytrade.shtml>.

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

Linux is a trademark of Linus Torvalds in the United States, other countries or both.

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 or other public sources. Questions on the capabilities of the non-IBM products should be addressed with those suppliers.

All performance information was determined in a controlled environment. Actual results may vary. Performance information is provided "AS IS" and no warranties or guarantees are expressed or implied by IBM

The IBM home page on the Internet can be found at: <http://www.ibm.com>.

