

WHITE PAPER

A New Strategic Approach To HPC: IBM's Blue Gene®

Sponsored by: IBM

Christopher G. Willard, Ph.D. Addison Snell

Earl Joseph, Ph. D.

December 2004

NEW STRATEGIES FOR HIGH-PERFORMANCE COMPUTING

IBM is experimenting with new ways to advance its position in the high-performance computing (HPC) market by applying a product strategy based on continuous market disruptions around new technologies that address specific customer requirements. IBM's Blue Gene® computer system represents the first product from this new strategy with the goal of setting a new direction for high-end computer designs.

To continue its success, IBM has taken a new strategic approach to the market that would address both some weakly served market niches and the problems related to ever-growing system sizes resulting from current product design approaches. In the future, metrics like: TFLOPS/watt, TFLOPS/sq ft will become more important than TFLOPS/\$ to datacenter managers due to the growing size of systems combined with the increase in power and heat with faster processor technologies. Price/performance will always be important, but may no longer be the dominant metric.

In developing Blue Gene, IBM system architects moved the focus away from high single-processor clock ranges and theoretical price/performance numbers, choosing instead to concentrate on the HPC requirements in these areas: high levels of system scalability, processor and system density, power consumption, system cooling, systems management, reliability and familiarity of programming environment. They did this so that buyers could both afford to acquire these large systems and to fit them into their operational environment without also investing heavily in expanding power capabilities, cooling, and adding new buildings.

We see this move as a strong indication that IBM designers believe they have been able to overcome some of the design issues in today's parallel computers, and we expect that this new design approach has the ability to effectively scale problems in order to obtain performance.

In the future metrics like: TFLOPS/watt, TFLOPS/sq ft will become more important than TFLOPS/\$.

IN THIS WHITE PAPER

This white paper examines the development of IBM's Blue Gene system, with emphasis on the company's HPC technology development strategy and its evolution through the project. More specifically the paper provides:

- ☒ An overview of IBM's approach for bringing HPC systems to market
- ☒ An overview of the Blue Gene system design with emphasis on unique architectural features.
- ☒ A review of IBM's product market strategy for Blue Gene
- ☒ IDC's analysis of the future, opportunities, and challenges for new capability systems such as Blue Gene

THE NEW STRATEGIC APPROACH

To expand on its success in the technical computing market, IBM has decided to launch a new strategic approach to the market that would address both some weakly served market niches and the problems related to ever-growing system sizes resulting from current product design approaches. In the future, metrics like: TFLOPS/watt, TFLOPS/sq ft will likely become more important than TFLOPS/\$ to datacenter managers. Price/performance will always be important, but may no longer be the dominant metric.

Long-term success in the computer industry requires vendors to continually update and enhance their product strategies to meet rapidly growing user requirements while incorporating advances in new technologies. Scientists and engineers using high-performance computers (a.k.a. supercomputers) are the most demanding "bleeding-edge" users of computers, and they are always demanding more from HPC vendors.

Technical computer users have an insatiable demand for computational performance, driven by the nature of their research and development. The scientific research and engineering development cycle is to first solve one problem, which in turn creates many new problems that are always more difficult than their predecessors, and thus require more powerful tools to solve, including more powerful computers. This group has problems that require 10x, 100x, and often 1,000x more computational power than what they have available at any given point in time.

Technical computer users have an insatiable demand for power, driven by the nature of their research and development.

Over the last decade, IBM has addressed HPC markets with a leveraged product strategy that provided strong price/performance by leveraging higher volume building blocks. This strategy worked well for the majority of the HPC market, but has not addressed certain segments that demand special, more advanced capabilities. This strategic approach has also led to system designs that are becoming very large with 1,000's of processors requiring extensive amounts of power, cooling, and floor space.

Market Experimentation

IBM hopes this new strategic approach will lead to technologies that can then be applied in other areas across IBM. In many ways this strategy is one of continuous market experimentation:

a strategy of
continuous market
experimentation

1. Take a serious look at market segments that are either poorly addressed by today's products or segments that will soon be mismatched with the next generation of "normal" product designs.
2. Use this knowledge to guide one through the internal research projects — select the most promising new ideas to pull out of the research labs — and design a new class of products around these technologies.
3. Quickly deliver the new products to the market and gauge the response.
4. Expand on the ones that stick and try again with the ones that don't.

Blue Gene is the first product that IBM is bringing to market under this new strategy. In developing Blue Gene, IBM system architects moved the focus away from high single-processor clock ranges and theoretical price/performance numbers, choosing instead to concentrate on the HPC requirements in these areas: high levels of system scalability, processor and system density, power consumption and system cooling. This approach resulted in improved reliability — by virtue of cooler operating components and fewer moving parts. They did this so that buyers could both afford to acquire these large systems and to fit them into their operational environment without also investing heavily in expanding power capabilities, cooling, and adding new buildings.

IBM's Enhanced HPC Computing Approach

IBM's development of its Blue Gene system represents a new approach to bringing a new class of capability systems to market. This approach has three major components:

- ☒ Investment in broad areas of new technology through IBM's internal research and development efforts. This leverages IBM's on going R&D investments.
- ☒ Collaboration with the scientific community to help define requirements, identify new opportunities, develop applications, help with system testing, etc. The emphasis here is to work with end users during the system design phase. In this regard, IBM collaborated with Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Sandia National Laboratory, Argonne National Laboratory, San Diego Supercomputer Center, Columbia University, and California Institute of Technology.
- ☒ Leveraging IBM's overall infrastructure for building the specific actual system designs. IBM has the ability to make use of a broad array of internal capabilities in developing and fielding new generations of technologies. These capabilities

range from component parts, to manufacturing, to system software, to services and support, to new approaches to accessing resources.

Blue Gene in the Early Years

When chartered five years (1999) ago the Blue Gene system was originally conceived as a one-of-a-kind, single application computer to address protein folding problems. As the development project gained momentum with the availability of operational hardware, the IBM team and its partners identified other classes of problems that the system could effectively address, and IBM moved to provide a more generally available system.

Interesting system design goals included: the use of low-power embedded processors, the development of system-on-chip ASICs, the development/use of dense packaging to minimize floor space requirements, providing a system with competitive cost/performance, and minimizing total power consumption.

New Design Rules Were Driven By Customers

The high-performance computing community is driving the vendor community to "change the design rules for system architectures" in support of their quest for Petaflops levels of computing performance. In order to accomplish this level of computational performance, computer manufacturers need to rewrite the design rulebook across a range of architectural characteristics:

- ☒ Increase effective scaling from a few thousand of processor to systems that run well with hundreds of thousands of processors. This requirement goes well beyond simply stacking a large number of CPUs in a room, to creating a coherent system that allows users to build programs that take advantage of large proportions of the available computer power.
- ☒ Create operationally "practical" systems that can fit within existing computer rooms, consume a reasonable amount of power, and require roughly the same amount of cooling as existing systems.
- ☒ Provide field manageable systems that allow existing systems operations staff to configure, monitor, upgrade, assign, and control system resources; as well as providing strong support for keeping both the system hardware and software up and running. Users wish to reduce investments in people to "run" computers.

Blue Gene Today

IBM responded to customer requirements by refocusing Blue Gene to meet a host of customer concerns regarding where current product designs are heading. This year IBM fielded operational systems for user evaluation and testing. The current design goals of Blue Gene aim for a scalable supercomputer having up to 65,536 dual-processor compute nodes and target peak performance of 360 teraflops.

Blue Gene Architecture Overview

IBM has designed its Blue Gene architecture for maximum affordable scalability on parallel applications. The Blue Gene architecture defines a massively parallel system based on single-chip "nodes" that are combined into larger systems through a hierarchy of form factors and several independent communications networks. To achieve maximum performance per square foot (and per watt), IBM traded off individual CPU performance for more scalability in a denser form factor.

The Building Blocks: Node-Level Architecture

The node contains two PowerPC 440 processors in a single-chip dense packaging design to maximize processing performance per chassis. The clock rate is relatively modest (700MHz) by design, both to increase density and to improve the ratio of memory latency-to-clock speed.

The floating-point units within a core are driven by the same instruction unit and thus operate as a SIMD unit or a vector processor with length 2 vectors. Each processor can perform 4 floating-point operations per cycle, for a total of 5.6GFlops per node.

Each node also carries a three-level cache hierarchy and 512MB of DDR memory in its current implementation. Each processor has a 32KB instruction cache and a 32KB non-coherent data cache integrated with the 440 core, as well as a small, dedicated 2KB L2 cache with a round-trip latency of 16ns (approximately 11 clock cycles). The L3 cache on the chip ASIC is shared by the two processors, as is the main memory.

Depending on the nature of the application to be run on Blue Gene, the programmer may choose to employ both processors in a node for computation, or have one processor dedicated as an I/O processor to handling message-passing operations. Dedicating one of the processors to MPI communications will halve the theoretical peak speed but will increase the available cache and memory to the computing processor and can significantly reduce MPI latency.

System-Level Architecture

The dual-processor node ASICs are connected over multiple independent interconnect hierarchies to create Terascale systems. With redundant paths, processors are connected in both torus and collective network topologies. Blue Gene systems are built up from components listed in Table 1.

TABLE 1
Blue Gene System Architecture Components
Compute Cards
<p>These cards are the basic building block and smallest field-replaceable unit in a Blue Gene system.</p> <p>Each compute card contains two of the dual-processor nodes described above, for a total of 4 processors (11.2GFlops).</p> <p>The total off-card I/O bandwidth is 2.8GB/s for the torus, and 2.8GB/s for the collective network.</p>
Node Cards
<p>There are 16 compute cards per node card.</p> <p>A single node card can be configured with 32 nodes, 64 processors, 16GB DDR memory, running at a peak speed of 180GFlops.</p> <p>The total off-card I/O bandwidth for torus is 22.4GB/s, and 2.8GB/s for the collective network.</p>
Racks
<p>A single rack can be configured with 1024 nodes, 2048 processors, 512GB DDR memory, running at a peak speed of 5.7Tflops.</p> <p>At the rack level, the total off-rack I/O bandwidth for torus is 224GB/s, and the peak bandwidth of collective network is 7GB/s, although IBM normally uses 4.2GB/s.</p>
Systems
<p>Multiple racks can be further combined into larger systems, extending the rack-level architecture.</p> <p>The maximum system size, ordered by Lawrence Livermore National Laboratory, is 64 racks, or 131,072 processor cores with a theoretical peak speed of over 360Tflops.</p>
Service Nodes
<p>Blue Gene also requires a service node where the system administrator manages the complex, front-end nodes, where end users compile and launch jobs, and file servers for storing data.</p>

Source: IDC, 2004

Flexibility In Design: System Partitioning

Blue Gene systems can be logically divided or partitioned into several smaller independent systems. Partitions maintain the same architectural characteristics — node hierarchy, interconnect topology, etc.— as the larger system. Partitioning allows the systems to be used by multiple users who are testing code or simply do not need the entire machine. It also allows operations staff to test new versions of system software in one partition while running current software versions on the rest of the system.

Interconnect Networks

The heart of multiprocessor computers is the interconnect. The interconnect is a defining feature for system architecture type (e.g., shared memory, distributed memory, NUMA); it sets bounds on system balance and is a determinant in performance of memory and I/O bound applications. For parallel computers the performance of the interconnect defines the type and size of applications that can be effectively run on a system. The more powerful the interconnect, in terms of latency and bandwidth, the greater the range of applications that can be addressed and the greater the scalability of a given application on the computer. Parallel computer interconnects support a range of functions, including node-to-node memory communications, program thread coordination and control, supporting global operations, I/O and system management communications.

Before looking at the specifics of the Blue Gene interconnect, it is important to take a brief look at communications costs. These costs are a factor in determining the ultimate scalability and performance of a parallel application. Communications can involve either data transfers or control signal-to-synchronized task activities. Communications overhead cost includes time spent creating messages, executing communications protocols, physically sending messages, and running through the protocol sets, and decoding the message on the receiving node. The effect of overhead delays is defined mathematically by Amdahl's Law, which can be summarized as follows: in computer operations with slow and fast components, the slow component ultimately determines the speed of the operation. Overhead is the most egregious type of a slow component with even low percentages of overhead limiting the effective scalability of an application. The Blue Gene system was designed with an extensive set of interconnect features to help reduce or eliminate sources of communications overhead.

IBM developed five separate interconnects for the Blue Gene system, each designed to address a specific set of functions, and all of which are implemented on the node chip (as shown in Table 2).

TABLE 2

Blue Gene: Five Separate Networks

#1: Three-Dimensional Torus

A three-dimensional torus (3D torus) network is used for point-to-point messaging between compute nodes.

A 3D torus architecture connects nodes along logical x-, y-, and z-axis grid (or up/down, north/south, east/west directions); the grid is made into a torus by linking the end of any grid line back to the start of the line.

The advantages of this interconnect architecture include an easily understood and regular logical arrangement of nodes, no special processing for edge nodes (the loop-back feature of a torus eliminates edges), the communications path lengths grow much more slowly than the number of nodes (path lengths scale linearly, while the number of nodes scale as a cube), and the ability to send messages in either direction around a ring effectively cuts in half the distance between farthest points.

#2: Global Collective Network

In technical applications, it can be necessary to perform operations across large data sets, such as summation of data or finding a maximum value. These collective operations can be expensive on distributed memory machines because they generate a large amount of message traffic for a small amount of computational work.

IBM addresses this issue by providing a second network with the ability to perform collective operations within the network itself.

From a technology perspective the important point here is that the Blue Gene collective network performs the operations within the network itself rather than requiring nodes to decode messages with intermediate values, calculate new intermediate values, create new messages, and send them on to other nodes. Most of the time spent in this latter case would be pure overhead associated with the communications protocols.

#3: Global Barrier and Interrupt Network

Parallel applications need to coordinate the work being done by the multiple tasks running in parallel. For example, a set of tasks may all need to complete before the applications can move on to a new phase (a barrier operation).

The amount of coordination varies significantly between types of applications and can be a limiting factor in what applications can be effectively parallelized.

Coordination is another overhead operation that can be very expensive to run through a message-passing process. Blue Gene has a separate communications network devoted to speeding up task-to-task coordination activities.

#4: Ethernet Network for Machine Control

The Blue Gene system includes a 100Mb Ethernet network for machine control.

This network provides direct access to all nodes, and is used for operations such as system boot, debug, and access to performance counters.

#5: Gigabit Ethernet External Interface Network and System I/O

A final area of computer communications is between the system itself and the outside world (i.e., input/output or I/O).

The Blue Gene systems use a Gigabit Ethernet network to communicate with all I/O devices and local area networks.

This network is run from specially designated I/O nodes that pass data to the rest of the system via the global tree network.

Environmental Factors

Environmental factors, such as power consumption, cooling, and floor space, are becoming critical to large HPC acquisitions. IBM has made technology investments in Blue Gene to make these costs scale to a different set of curves compared to today's computer designs.

The race in the industry for higher theoretical gigahertz ratings has increased the number of transistors on a processor, which in turn increases the power consumption and heat dissipation. Supercomputers based on processors with high power consumption and heat profiles usually require additional space for air flow through and around components and often require additional expensive, exotic cooling systems to maintain operations.

The Blue Gene architecture is designed for lower power consumption and higher computational density. IBM's approach is to obtain processor-level performance by managing the ratio of memory and I/O access to compute time. This approach aims to provide equivalent performance at lower power levels to what could be obtained by simply ratcheting up the clock on the processor. By managing the power consumption, IBM fits 2,048 processor cores into each Blue Gene rack.

Between racks, IBM employs a simple innovation to improve cooling efficiency. Supercomputer labs are frequently organized with rows of racks facing in opposite directions, such that the racks will be face-to-face in one aisle, then back-to-back in the next, alternating across the room. In these layouts, there will be alternating "hot" (exhaust) and "cold" (intake) aisles. Blue Gene manages air flow for cooling by putting a diagonally partitioned plenum between racks. This technique matches the volume of air in the plenum to the amount of airflow needed to cool the rack. For example, the amount of cold air coming in from the floor is at its highest near the bottom of the rack. As it travels up the rack, the amount of airflow needed decreases, the reduction in volume from the slanted baffle matches the amount needed near the top of the rack and for the exhaust of hot air.

Systems Software

To develop an operating system for a parallel computer is the same level of effort as to describe it in detail briefly. — Ancient Cybernetic Proverb.

There are two major and sometimes conflicting requirements for systems software on parallel computers, the first is to provide a fully functional and easily used system management and programming environment. The second is to minimize system software overhead on computational nodes.

Blue Gene runs a specialized lightweight O/S kernel on compute nodes, Linux on I/O nodes, and on front-end and service nodes. The compute node kernel performs basic resource management functions and is designed to minimize O/S overhead costs, in particular, to eliminate the interference from various service processes.

Blue Gene provides a standard Linux distribution running on the systems front-end nodes (those which users see). These nodes are typically standard pSeries machines and are connected to Blue Gene over the Gigabit Ethernet. There may be one or more front-end nodes depending on the number of users of the system. System administrators manage Blue Gene complexes from a service node running Core Management and Control System (CMCS). The key functions of CMCS are system configuration, initialization, monitoring, and operation.) In addition, IBM plans to provide a number of standard and specialized system software tools, including:

- ☒ **LoadLeveler.** Manages job submission and workload balancing. LoadLeveler will coordinate with a Blue Gene-specific scheduler function that selects a set of compute nodes to form a partition that meets the size and shape requirements for a given job as specified by the user.
- ☒ **IBM General Parallel File System (GPFS).** GPFS is a shared-disk file system that will provide data access from all nodes in a Blue Gene complex. Applications can access shared files using standard file system interfaces. In addition, a single file can be accessed concurrently from multiple nodes.

The Blue Gene applications development environment includes:

- ☒ **Compilers.** Standard IBM XL Fortran, C, and C++ compilers for PowerPC. These compilers are augmented with a back end that takes advantage of the dual floating-point units on Blue Gene nodes.
- ☒ **IBM Engineering and Scientific Subroutine Library (ESSL).** A collection of over 400 mathematical subroutines for applications written in FORTRAN, C, or C++. IBM is tuning routines in the library for the Blue Gene architecture.
- ☒ **Message-passing support.** IBM provides the Message Passing Interface (MPI) library to parallel applications support. This implementation is based on the MPICH2 library from Argonne National Laboratory. It supports low-latency communication (the latency between MPI processes on neighboring nodes is 3.3 μ s) and exploits the global collective network, the global barrier, and interrupt network, and the multicast feature of the torus network for certain collective communication operations, such as: all reduce, broadcast, and barrier synchronization.

FUTURE OUTLOOK

There are three basic approaches to developing HPC computers. The first uses standard commodity server and networking products to produce a low-cost cluster. Clusters have proven to be cost effective in many throughput computing environments and effective capability machines for a set of parallel applications. The second, approach is to design custom processor and/or components and software as in traditional vector supercomputer designs. These computers often provide the highest level of performance but come with a higher price tag. A third approach is to design systems specifically to run parallel applications using standards and commodity technology where appropriate but incorporating specialized technology when needed. These purpose-built systems have historically extended the reach of capability computing by both increasing the performance of traditional parallel applications and increasing the number of applications that can be effectively parallelized on the computer. In addition these systems have often proved to be effective throughput-oriented capacity machines. The Blue Gene system fits into this latter category, but with a number of design innovations.

One of the unique features of the Blue Gene system is not just its architecture, but the investment in its development. IBM is one of the few companies in the world that is able and willing to invest in the development of new system architectures and that has the internal resources to bring such architectures out of the lab and into the market. The company has the advantage of being able to bring a combination of internal R&D resources, previously developed components, and internal manufacturing capabilities to bear in the development process. That said, the development of any new system represents a large financial outlay — usually in multiples of hundreds of millions of dollars.

Although all research efforts do not have to become commercial successes, the move to bring a product out of the lab and into the market (even in limited form) incurs additional direct opportunity costs for the company. We believe that IBM is expecting return on its investment in Blue Gene in a number of areas: First, the product can help position IBM as a premier contender in the HPC market and thus support sales of the company's overall technical computing line. Second, if IBM is able to develop an expanded product line that addresses a significant portion of the market, then it can expect direct financial return from sales and services. Third, much in the same way IBM's RS6000 SP product line started life as a technical computer and ended its career in predominantly business-oriented markets, the Blue Gene architecture might be expanded to include commercial versions of the product. Finally, technology and concepts developed and proven in the course of creating and bringing the system to market may find their way into future company products. IBM is also providing on-demand capability, which enhances end-user access to technology.

Challenges

The major challenges for IBM with the Blue Gene product fall into three broad areas: managing complexity, developing an applications base, and extending product reach.

- ☒ **Complexity.** The Blue Gene system with its multiple layers of system scaling and multiple network types appear to IDC as a complex architecture even by HPC standards. Complexity in and of itself is not bad, particularly when there are clear reasons for different features. That said, IBM must demonstrate that the computer does function efficiently as a system without internal conflicts or unexpected behaviors. In addition, the product must be programmable by a broad cross section of the scientific and engineering community, this requires both internal mechanisms to hide complexity, strong optimizing compilers, and a complete programming tool set.
- ☒ **Developing an applications base.** Although supporting the programmability of a system is a major requirement for system acceptance it is only a first step. Long-term success of technical computing systems depends in large part on the vendor's and user's abilities to convince third-party software developers in both the private and public sectors to port and optimize their applications to the new system. This is generally a long-term process, with vendors supporting the development of an applications library through direct technical help, grants, and contractual agreements with software vendors.
- ☒ **Extending the product reach.** Ultimately the success of Blue Gene will depend on the extent that Blue Gene base products or technologies appear throughout the market. For technical markets the company needs to develop products that match a broad set of price points and that can be used in both capability and capacity modes. For commercial markets IBM can either work to identify applications areas that a Blue Gene derivative would address and/or propagate the technologies and concepts through future product lines.

CONCLUSION

IDC believes that at least at the high-end of the market, a new technology cycle is beginning, with users moving away from amassing large numbers of low-cost computer cycles and toward supporting efforts to develop "break through" technologies that will allow them to effectively address their next-generation problems.

The HPC technical server market landscape is evolving as disruptive technologies continue to redefine the market space. First leveraged technologies provided cost-effective solutions by leveraging moderately high volumes in the computer's components, then clusters based on considerably higher volume commodity building blocks redefined price/performance, what will be the next wave of change? The next wave will have to address:

IDC believes that at least at the high-end of the market, a new technology cycle is beginning.

- ☒ Increased scaling to very large processor counts and large memory sizes
- ☒ The need to power and cool Peta class systems without adding major additional investments
- ☒ The ability to fit Peta class systems within standard computer rooms, as only a few sites can afford to construct new dedicated buildings
- ☒ How to use and manage systems that are 10x and 100x larger and more complex

An interesting aspect of the Blue Gene system is that it uses large numbers of smaller lower-power processors in order to reduce power and cooling demands while providing very high computational densities. We see this move as strong indication that IBM designers believe they have been able to overcome some of the design issues in today's parallel computers and expect that the system has the ability to effectively scale problems in order to obtain performance.

IBM plans to leverage its strategic investments in Blue Gene across other server segments as the technology matures in the technical server market. Commercial datacenters will be faced with the same environmental constraints that technical data centers are just starting to experience, only at a later point in time. The innovative elements of the Blue Gene architecture could find their way into IBM's next-generation server designs, thereby making Blue Gene a more pervasive design approach for all large IBM servers.

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 2004 IDC. Reproduction without written permission is completely forbidden.