Arquitetura IBM Blue Gene/L para Supercomputadores

Maíra Saboia da Silva
RA:098338
MO801 - Arquitetura de Computadores I
Campinas, SP – Brasil
mairasaboia@gmail.com

RESUMO

O projeto Blue Gene trouxe grandes inovações para as arquiteturas de supercomputadores. Neste trabalho é apresentado um resumo da evolução história, finalidade e uma visão geral sobre a arquitetura Blue Gene/L, primeiro supercomputador da sério dos Blue Genes.

Categories and Subject Descriptors

C.5.1 [**Computer System Implementation**]: Large and Medium ("Mainframe") Computers – Super (verylarge) computers

General Terms

Desenmplenho, Projeto de Hardware, Experimento.

Keywords

Supercomputadores, arquitetura, Blue Gene, Blue Gene/L, alto desempenho, hardware

1. INTRODUÇÃO

O blue Gene/L - (BG/L) é uma arquitetura de computadores projetada para produzir supercomputadores que processem dados a uma velocidade de PFLOPS (petaFLOPS - Quatrilhão de Operações em Ponto Flutuante por Segundo). Anunciada em 2001, ela surgiu de uma parceria entre a IBM, o *Lawrence Livermore National Laboratory - LLNL* e o *United States Department of Energy*[1].

Esse projeto foi uma iniciativa para construir um supercomputador que explorasse as fronteiras da supercomputação: na arquitetura do computador, no software necessário para controlar sistemas massivamente paralelos e no uso da computação para promover o avanço no entendimento dos mecanismos moleculares via simulação em larga escala.

O Blue Gene/L foi o primeiro supercomputador da série Blue Gene, constituída de quatro supercomputadores: Blue Gene/L, Blue Gene/C, Blue Gene/P e Blue Gene/Q[1].

O termo *Blue*, de seu nome, foi escolhido por corresponder à cor utilizada pela IBM e o termo *Gene* representa a finalidade para a qual o Blue Gene foi projetado. Na literatura, seu nome algumas vezes se refere ao computador instalado no LLNL e outras vezes se refere à arquitetura de tal computador. Neste texto, Blue Gene/L faz referência à arquitetura desse computador.

Em 2004 um protótipo do Blue Gene/L superou o supercomputador mais rápido naquele momento, o *Earth Simulator's*, cujo desempenho era de 35.86 TFLOPS (teraFLOPS) no *benchmark* HPL, atingindo 36.01 TFLOPS - dados divulgado pelo projeto TOP500, que ranqueia e detalha os 500 (não distribuídos) mais poderosos sistemas computacionais conhecidos no mundo.

Posteriormente, o Blue Gene/L foi o primeiro supercomputador a executar mais de 100 TFLOPS sustentado em uma aplicação real do mundo. Com isso, ele ganhou o 2005 *Gordon Bell Prize*.

Ocupou a primeira posição até 2008, perdendo-a para o *Roadrunner*, também projetado pela IBM, que alcançou o desempenho de 1.026 PFLOPS. Nesse momento, o Blue Gene/L tinha desempenho de 478.2 TFLOPS (petaFLOPS – Milhares de Trilhões de Operações em Ponto Flutuante por Segundo). Em junho de 2010, mantendo o mesmo desempenho, o Blue Gene/L caiu para a oitava posição. O *Jaguar* vendido pela *CrayInc.* é atualmente o supercomputador mais rápido do mundo, seu desempenho alcança 1.75 PFLOPS executando no *benchmark* do Linpack, todavia, sua capacidade teórica é de 2.4 PFLOPS.

2. VISÃO GERAL DO BLUE GENE/L

O principal objetivo da supercomputação é criar computadores que executem programas o mais rápido possível. A forma típica de promover a supercomputação é agrupando um conjunto dos mais rápidos computadores que a tecnologia possa permitir e deliberar para cada um deles uma quatidade de responsabilidades de computação. Contudo, os poderosos multiprocessadores simétricos (*symmetric multiprocessor*- SMP) normalmente utilizados

em supercomputadores consomem uma quantidade grande de energia, além de que a taxa de processamento não aumenta à mesma taxa que o aumento do núumero de processadores. Por conseguinte, os projetistas do Blue Gene/L utilizaram uma abordagem diferente. O supercomputador foi implementado agrupando-se um "grande número de nós" cada um com uma taxa de clock modesta de aproximadamente 700 MHZ. Esses nós são eficientes e de baixo custo, viabilizando o agrupamento milhares deles. Cada nó é responsável pela realização de tarefas simples realizadas paralelamente. Esse conceito permite que o Blue Gene/L possa processar eficientemente grandes quantidades de dados[2].

O projeto do Blue Gene/L tem os seguintes aspéctos:

- Processador IBM PowerPC 440
- DRAM embutida
- Tecnologia *System-on-a-Chip*
- Projeto de processador dual

O projeto de processador dual utiliza um processador para computação e o outro para comunicação.

2.1 Estrutura do Hardware

Para implementar essa abordagem é necessário agrupar muitos nós. O sistema BG/L é escalável, podendo atingir um máximo de 2^{16} (65536) nós de computação, configurados como uma rede *Torus 3D* com dimensões de 64x32x32. O Blue Gene/L é construído usando a tecnologia *System-on-a-chip*, que significa que todas as funcionalidades de cada nó, com exceção da memória, estão dentro de um único circuito integrado ASIC (*Application Specific Integrated Circuit*). Esse chip inclue dois núcleos PowerPC 440 de 32 bits, que foi desenvolvido pela IBM para aplicações embarcadas e, cada nó possui 2 gigabytes de memória local. Associado a cada nó tem uma dupla Unidade de Ponto Flutuante (*Float Point Unit* - FPU) que opera no modo SIMD (*Single Instruction, Multiple Data*) [3].

Sua configuração física é caracterizada por uma estrutura modular de 5 níveis. O componente básico é o chip ASIC que contém dois processadores e é denominado de nó de computação; os nós ASICS são armazenados de forma que dois nós estejam agrupados em um *Computer Cards*; um *Node Board* ou *Node Card* armazena 16 *compute cards*; um conjunto de 16 *node card* podem ser armazenado por um *Midplane* e uma *hack* suporta dois *Midplane*. Além disso, com uma configuração final de 64 hacks agrupados para formar um sistema é possível alcançar 65536 nós de computação. A Figura 1 representa a estrutura modular projetada para o Blue Gene/L[4].

Como cada *hack* armazena dois midplanes, é possível armazenas 1024 nós por *hack*. Quanto mais *hack* são adicionados, o sistema é capaz de alcançar maior velocidade. Cada processador pode executar cerca de 4 operações de ponto flutuante por ciclo, o que leva a um desempenho de cerca de 1,4 teraflops num único *midplane*. Teoricamente, cada *hack* pode acrescentar cerca de 2.8 teraFLOPS de velocidade de computação. Na prática, os números são sensivelmente menos do que isso.

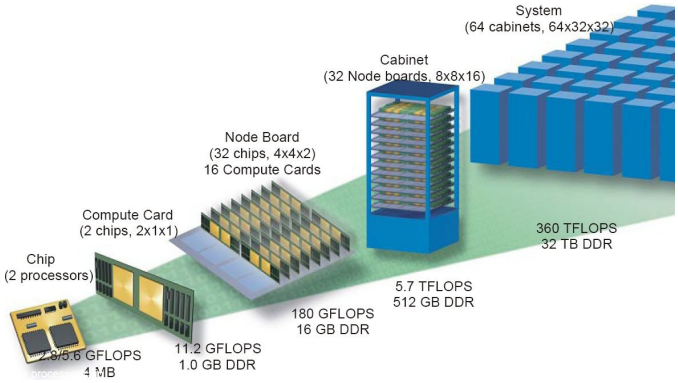

Figura 1-Estrutura Blue Gene/L

Quando BG/L superou o Earth Simulator como o computador mais rápido, alcançou uma velocidade de 36,01 teraflops. O número de racks mais tarde duplicou, atingindo uma velocidade de quase 70,72 teraflops. Mais recentemente, quando o número de racks foi novamente dobrado para 32, o BG/L alcançou 135,5 teraflops, quase o dobro do seu recorde anterior.

Além dos 65.536 nós de computação, o sistema tem 1.024 processadores de I/O. Cerca de um nó de I/O para 64 nós de computação. Os nós de I/O possuem o mesmo ASIC dos nós de computação, mas com uma memória externa expandida. Cada nó tem um pequeno kernel que manipula comunicação e funções básicas que podem aumentar o desempenho necessário nos programas científicos. Os nós são interconectados usando cinco redes diferentes: torus, rede coletiva (tree) , Ethernet, JTAG e interrupção global. A principal comunicação ponto-a-ponto na rede para mensagens é realizada pela rede torus 3D.

As duas redes mais importantes para comunicação são a rede torus e a rede tree. O BG/L faz sua comunicação geral via a rede torus. A rede torus conecta todos os nós, dando a cada um deles 6 nós adjacentes. Um nó é posicionado à direita, um à esquerda, um à frente, um atrás, um em cima e um em baixo. A largura de banda para essas ligações são de 2 bits/ciclo ou 175MB/s a uma taxa de 700 MHz.

Cada mensagem transmitida é fragmentada em diferentes pacotes que possuem tamanho num intervalo de 32 a 256 bytes em frações de 32 bytes. O BG/L usa uma rede *tree* para comunicação coletiva que ocorre com muita frequência. Uma rede tree é uma rede que combina dois ou mais redes *star*. Redes star são redes onde todos os nós das estações de trabalho são ligadas ao nó central. Na rede tree do Blue Gene/L, cada um desses nós centrais *star* são ligados juntos. A rede BG/L tem uma largura de transmissão de 350 MB/s, muito mais rápida do que o link da rede de torus[5].

2.2 Componentes do Sistema

Nesta seção estão descritos alguns componentes que pertecem aos sistema Blue Gene/L.

2.2.1 Nós

Cada nó do Blue Gene/L, tanto de computação quanto de I/O, é baseado no processador disponibilizado pela IBM. Esse nó contém:

- Uma dupla unidade de ponto flutuante por núcleo do processador;
- Caches individuais de dados e instruções (32 KB)
- Uma pequena cache L2 que serve principalmente como buffer de pré busca.
- Uma cache L3 compartilhada de 4 Megabit contruída com tecnologia DRAM embarcada;
- Um controlador externo de memória DDR integrado;
- Um adaptador Ethernet Gigabit

A unidade de ponto flutuante (*Floating Point Unit* - FPU) é uma unidade de execução dedicada que opera nos números de ponto flutuante. O Blue Gene/L tem uma unidade dupla de ponto flutuante (*Doule Floating Point Unit* - DFPU) em sua arquitetura alcançada através da junção de duas FPU, uma é a unidade primária e a outra a secundária. A FPU secundária é basicamente um duplicado da primeira, mas a segunda possui um conjunto especial de funções paralelas. Todas as instruções SIMD (Single Instruction, Multiple Data) são realizadas em dados de ponto flutuante. Mas isso não significa que a segunda FPU não possa realizar operações de ponto flutuante [6].

Geração de código para DFPU pode ser feita com um compilador IBM XL TOBEY. Ele traduz códigos fontes C, C++ ou FORTRAN para linguagem intermediaria. O compilador TOBEY reconhece onde ocorrem as computações de ponto flutuante e então usa extensões do BG/L para implementar eficientemente as computações.

2.2.1.1 Nós de Computação

O nó de computação é a unidade básica de computação. Ele consiste de dois circuitos ASCIs e memória SDRAM (*Synchronous Dynamic Random Access Memory*) de 512 MB ou de 1 GB. Cada processador executa seu próprio kernel de computação, que é baseado em um Linux otimizado para o ambiente do Blue Gene/L. Nós de computação não executam diretamente operações de entrada e saída, mas isso é invisível para o usuário. Os programas executam operações de entrada e saída de forma usual, e o kernel do nó de computação transfere a requisição para o nó de entrada e saída responsável.

A medida que o processador PowerPC 440 não implementa o hardware necessário para prover suporte ao SMP, os dois núcleos não possuem caches L1 coerentes. Todavia, as caches L2 e L3 são coerentes entre os núcleos dos nós[6].

2.2.1.2 Nó de Entrada e Saída

O nó de entrada e saída (I/O - Input/output) é semelhante ao nó de computação, mas apresenta basicamente duas diferenças. A primeira é que o nó de I/O não está anexado à rede torus. A segunda diferença é que os nós de I/O sempre tem interfaces Ethernet.

O nó de I/O utiliza o mesmo chip ASIC que o nó de computação, mas tem memória externa expandida. Este nó estabelece comunicação entre os nós de computação e os dispositivos na rede funcional.

Cada nó de entrada e saída serve a um grupo específico de nós de computação, com uma proporção mínima de um nó I/O por 128 nós de computação. O grupo de nós de computação atribuído ao nó de I/O, é considerado um conjunto de processadores (pset). Cada nó de computação se comunica com o seu nó de I/O através de uma rede coletiva e os nós de I/O se comunicam com o mundo exterior através da rede funcional, que é projetada para ser uma rede Ethernet Gigabit. A figura 2 ilustra um mapeamento entre nós de computação e nós de I/O.

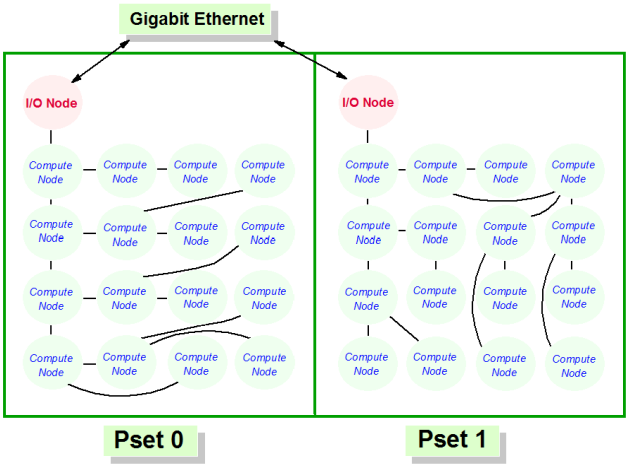

Figura 2 - Mapeamento pset[6]

2.2.1.3 Modos de Programação dos nós

Pelo fato das aplicações no Blue Gene/L serem executadas em paralelo, uma mensagem que indica o modo de programação é utilizada. O processador tem dois modos de execução: modo de comunicação e modo virtual. O modo padrão é o modo de comunicação, chamado de modo *Communication Coprocessor*, onde um dos processadores executa a aplicação e o outro é dedicado à transmissão de mensagem através da interface MPI (*Message Passing Interface*). O segundo é chamado de modo virtual, onde cada processador age como sendo independente um do outro, e executa sua própria cópia da aplicação do usuário. No entanto, contrariamente ao esperado, não é atingindo o dobro do desempenho, principalmente porque cada processador precisa administrar suas próprias operações de entrada e saída.

2.2.2 Node Card

O *Node Card* contém 16 nós nós de computação, e se conecta a 32 ASICs do tipo do Blue Gene/L. Além disso, um ou dois nós de I/O podem ser ligados a este cartão, que dá acesso ao disco do Blue Gene/L.

2.2.3 Midplane

Cada midplane contém quatro *link cards* que aceitam os cabos necessários para conectar midplanes juntos (e, portanto, as cabines). O Midplane tem espaço para 16 Node Card, cada um com 32 nós de computação ASICs. Cada midplane contém um *Service Card*, que distribui o clock do sistema e fornece função de controle para outras cabines.

2.2.4 Blue Gene/L cabinet

A cabine do Blue Gene/L é baseada em um rack disponível comercialmente. É normalmente projetado para acomodar dois midplane e permitir o fluxo de ar horizontal da esquerda para a direita necessários para esfriar aproximadamente 25 kW gerados por um rack completamente carregado com 1024 processadores.

3. HARDWARE DE COMUNICAÇÃO

Computadores escaláveis, como o Blue Gene/L, são compostos de muitos processadores que se comunicam um com o outro através de redes de interconexão. Muita atenção é dada à arquitetura de redes porque seu custo e desempenho afetam drasticamente a capacidade e viabilidade prática da máquina[7,8].

Há muitas razões diferentes para que os processadores se comuniquem, como passagem de mensagem ao nível da aplicação, sistema de arquivos de entrada e saída, e manutenção da máquina operando em nível de sistema. Normalmente, esses tipos de comunicação têm características individuais que podem ser exploradas para alcançar o máximo de desempenho por rede. No entanto, os custos práticos e limitações de espaço estabelecem como é possível compartilhar o menor número de redes que possibilitem todos os tipos de comunicação.

A comunicação do Blue é constituída através de cinco tipos de redes:

- A rede *torus* 3D para troca de mensagens ponto-a-ponto entre os nós;
- Rede Coletiva para as operações relativas a todos os nós da Rede;
- Rede de interrupção e restrição;
- Rede Gigabit Ethernet (JTAG) para controle de máquina;
- A segunda rede Gigabit Ethernet para conexão com outros sistemas, tais como servidores e sistemas de arquivos.

3.1 Redes do Blue Gene

3.1.1 Rede Torus

A rede Torus é uma das principais redes de comunicação na arquitetura do Blue Gene/L. Na rede torus, cada processador está conectado a outros seis processadores: dois na dimensão X, dois na dimensão Y e dois na dimensão Z. A figura 3 ilustra essa distribuição. Então, cada chip ASIC tem seis vizinhos mais próximos conectados. A largura de banda dos links da rede torus é de 175 MB por segundo em cada direção, totalizando 2.1 GB por segundo em cada nó.

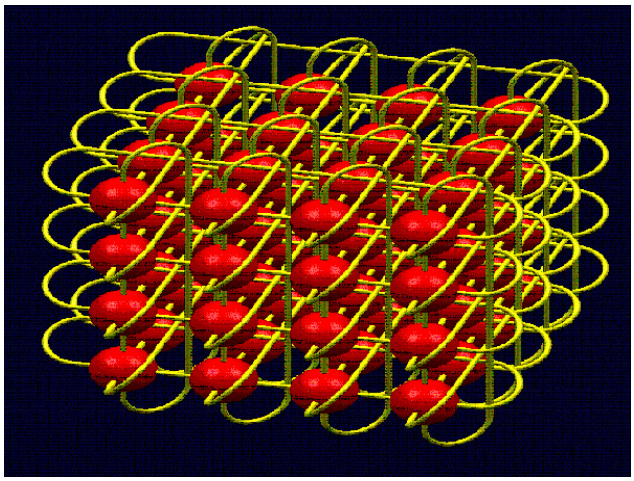

Figura 3 - Rede Torus 3D[6]

A rede torus é utilizada para propósito geral, troca de mensagens ponto-a-ponto e operações de *multcast.*

O clock interno do torus é um quarto da taxa do processador, então para 700 MHz, cada link do torus é de 175 MB por segundo. Existe barramento interno suficiente para os seis links de saída e de entrada trabalharem simutaneamente. Então, cada nó pode enviar e receber 1.05 GB por segundo. Ou seja, a largura efetiva de banda é de 175 MBps por segundo em cada direção.

3.1.2 Rede Coletiva

Em computadores escalares, é impraticável prover conexão entre cada processador e o sistema externo de arquivos. Normalmente, apenas um subconjunto de processadores ou processadores dedicados possuem ligações externas, e estas são compartilhadas entre todos os processadores.

Uma forma efetiva de alcançar esse compartilhamento é dividir os processadores e atribuir um subconjunto distinto a cada conexão externa. Para essa rede coletiva, o nó externo é chamado de raiz e prove conexão externa aos outros processadores do subconjunto dos quais ele é responsável.

Essa rede é utilizada para funções globais sobre todo o sistema Blue Gene/L. Cada processador de computação pode enviar mensagens para cima ou para baixo na rede e as mensagens podem ser destinadas a qualquer nível.

Os nós de computação também fazem parte da rede coletiva, mas não participam das funções globais. Todos os nós de computação existentes estão associados a um nó de I/O.

O tempo para viajar para cima e para baixo na rede colectiva é uma função da largura de banda do barramento, o número de portas, e os atrasos do pipeline através do hardware. Com um clock do sistema igual à frequência do processador, cada porta do barramento transfere quatro bits por porta em cada direção por ciclo de processador. A latência é baixa o suficiente para ser extremamente útil para as funções globais e para transferências entre os nós de computação e de I/O.

3.1.3 Rede Global de Interrupções e Restrições

Interrupção global é outra rede de inteconexão no Blue Gene/L. Ela fornece a funcionalidade necessária de obstáculos ou interrupções. Sua funcionalidade está intimamente relacionada com a funcionalidade das redes coletivas do Blue Gene/L e pode ser usada junto com ela para a passagem de mensagens.

A rede global de interrupção pode ser usada com o propósito de interrupções de software e restrições.

3.1.4 JTAG serviço de rede

Cada chip ASIC do Blue Gene/L, incluindo os nós de computação e nós de I/O, tem um endereço único que define sua localização na máquina Blue Gene/L. A JTAG (*Joint Test Access Group*) fornece esse tipo de serviço de rede e controle de máquina. Por exemplo, um nó de computação ASIC é definido por rack (0-N), midplane (0-1), node card (0-F), a posição do computer card em um node card (0-F), e sua posição no computer card (0-1). Este endereço é utilizado para acesso pelo barramento de serviço.

3.1.5 Barramento Gigabit Ethernet

Um grande Switch Gigabit Ethernet é o principal caminho de comunicação do Blue Gene/L com o mundo exterior. Este switch foi projetado para fornecer conectividade de alta velocidade para o CWFS (*Cluster-Wide File System*), que é o principal disco de armazenamento do Blue Gene/L. Esse switch fornece recursos de acesso a arquivos do sistema de arquivos. O barramento Gigabit Ethernet permite que todos os nós de I/O sejam acessados exclusivamente a partir do host do sistema. Este barramento é utilizado, por exemplo, como um checkpoint pelo sistema do Blue Gene / L para o disco.

4. SISTEMA DE SOFTWARE

Os princípios seguidos pelo software do Blue Gene/L são: atender os requisitos de escalabilidade e complexidade e disponibilizar um ambiente de desenvolvimento confiável.

O Blue Gene/L executa um Sistema Operacional personalizado. O SO controla o BlueGene/L, mais especificamente cada nó de computação e nó de I/O, apresentando características diferentes em relação aos sistemas operacionais convencionais. O nó de computação é gerenciado pelo sistema operacional CNK (Compute Node Kernel) e o nó de I/O pelo MCP (Mini-Control Program).

O CNK é um Kernel de propriedade da IBM, que implementa um subconjunto de chamadas de sistema do Linux. A maior parte destas instruções são para realização

de I/O como abertura e fechamento, leitura e escrita de arquivos, entre outras operações. O Kernel é monousuário, monotarefa e não possui mecanismo de paginação e é dedicado totalmente à aplicação sendo executada. O MCP é um Linux GPL que possui *patches* especiais para o Blue Gene/L.

A arquitetura de software do BG/L é composta por três entidades: um *núcleo computacional*, uma *infraestrutura de controle* e uma *infraestrutura de serviço*. Os nós de I/O fazem parte da infraestutura de controle. O código de usuário é rodado nos nós que formam o núcleo computacional. A infraestutura de serviços é feita por programas comerciais aparte do núcleo, conectadas ao resto da arquitetura por uma rede ethernet.

5. CONCLUSÃO

O projeto do Blue Gene/L explorou paradgmas de computação pudessem criar um computador de alto desempenho sem a utilização de processadores extremamente potentes e paralelos, mas utilizando grande quantidade processadores de baixo custo e baixo consumo de energia. Abordagem esta que não tinha sido utilizada até o momento da sua proposta. Com isso ele conseguiu ser o computador mais rápido do mundo liderando os rank TOP500 por alguns anos comprovando a viabilidade do projeto.

Também foi o primeiro computador a executar mais de 100 TFLOPS em uma aplicação. Com isso, ele ganhou o *Gordon Bell Prize em* 2005.

6. REFERENCES

[1] TOP500 Supercomputer Sites – Top List November 2010 [online].http://www.top500.org/list/2004/11/100, acessado em Junho de 2010.

[2] C. D. Gara A., Blumrich M.A. and C. G.L.-T. Overview of the blue gene/L system architecture. IBM Journal of Research and Development, Maio 2005.

[3] Milano, J.; Mullen-Schultz ,G.; Lakner, G.; *Blue Gene/L: Hardware Overview and Planning,* IBM Journal of Research and Development, Agosto 2006.

[4] Almási, G.; Bellofatto, R.; et. al. “An Overview of the Blue Gene/L System Software Organization”. Lecture Notes in Computer Science, EuroPar 2003, vol 2790, pp. 543555, Jun 2004.

[5] Almási, G.; Chatterjee, S.; et. al. *Unlocking the Performance of the BlueGene/L Supercomputer*. IEEE/ACM SC04, Pittsburgh, PA, Novembro 2004.

[6] Lakner G, *IBM System Blue Gene Solution: System Administration* - International Technical Support Organization - June 2007.

[7] H. Yu, R. K. Sahoo, C. Howson, G. Almási, J. G. Casta.nos, M. Gupta. *High Performance File I/O for The Blue Gene/L Supercomputer,* High-Performance Computer Architecture, International Symposium on, IEEE Computer Society, USA 2006.

[8] David Gregg and Jake Johnson , *The Blue Gene/L Supercomputer: Architecture and Implementation* . Technical Report, 2005.