Uma introdução ao IBM Blue Gene/L

Flavio C Buccianti
RA 075734
UNICAMP
IC - Instituto de Computação
www.ic.unicamp.br
flaviocb@gmail.com

INTRODUÇÃO

Este paper foi escrito com o intuito de dar uma visão geral sobre o IBM Blue Gene/L (BG/L) [1] desde o ponto de vista de arquitetura até a usabilidade e versatilidade para implmentação em aplicações de cálculos intensos e científicas.

Em 08/11/2004 o BG/L foi coroado como o computador mais rápido do mundo na 24a lista do TOP500 [2] que é um ranking dos 500 computadores mais rápidos do mundo e disponível em: www.top500.org. De 2004 em diante outros sistemas da familia BG/L de porte menor, passaram também a integrar a lista dos TOP 500.

Durante a fase de desenho do projeto do BG/L as preocupações principais foram com o alto nível de integração e baixo consumo quando comparado com máquinas da mesma categoria. Para poder satisfazer esses dois requerimentos foram utilizados os processadores embarcados PowerPC [4] System-on-a-chip (SOC)

Categorias e Descritores de Assunto (ACM)

C5.1 Large and Medium (“Mainframe”) computers

Termos

Design, Perfomance, Reliability

Palavras-chave

Blue Gene/L, TOP500, superocomputers, PowerPC

1.INTRODUÇÃO

O Blue Gene/L é o primeiro computador da família Blue Gene da IBM.

Em 2004 o BG/L entrou na lista dos computadores mais poderosos do mundo pela primeira vez, e já encabeçando a lista em primeiro lugar como o computador mais poderoso da face da terra. Tratava-se de uma máquina constituida de 16 racks contendo 16.384 nós de processamento com um indice de LINPACK [3] estável e mantido de 70.72 Teraflops (Trilhão de operações de ponto flutuante por segundo) e que representa 77% da sua performance de pico de 91.75 Teraflops. O BG/L completo é constituido, porém não está limitado a 64 racks contendo 65.536 compute nodes, ou seja 4 vezes o tamanho do BG/L que entrou no topo da lista do TOP500.

Cada node é constituido de 2 processador PowerPC 440 de 700 MHz com dois cores e 512 MB de memŕia SDRAM-DDR sendo encabeçados por um I/O node a cada 64 processadores ou compute node. O I/O node utiliza o mesmo chip do compute node. O sistema completo contém 33 Terabytes de memória.

Uma rede toroidal (Torus) tri-dimensional e uma rede coletiva (collective network) são usadas para interconectar todos os nós.

Desenhado para atingir 183.5 Teraflops como performance de pico usando um dos processadores de cada nó para computação e o outro para comjunicação. Quando se usa os dois processadores para computação, o BG/L atinge 367 teraflops de pico.

1.1Desenho do Blue Gene/L

Um diferencial do BG/L é o seu desenho de baixo consumo. Para atingir essa meta foi escolhido um processador eficiente em termos de consumo de energia: o PowerPC embarcado de 700 MHz.

Uma das premissas do projeto BG/L foi a de ter um alto nível de integração para simplificar os Racks. Para chegar à um alto nível de integração, a máquina foi desenvolvida com processadores com uma frequencia moderada baseado na tecnologia system-on-a-chip (SOC). Dentre as várias razões para escolher o design baseado em SOC estão: alto nível de integração, baixo consumo e baixo custo de projeto.

A escolha de um processador com uma performace modesta deve-se à vantagem que se tem em termos de performace/consumo desses processadores.

Baixo consumo é um dos pontos principais da família Blure Gene. A realção pode ser descrita por:

$$\frac{performance}{rack} = \frac{performance}{watt} \times \frac{watt}{rack}$$

onde watt/rack é determinado pela capacidade de resfriamento e pode ser considerado uma constante de 20 kW para um rack refrigerado à ar. É o termo performance/watt que determina a performance do rack. Isso ilustra claramente uma das áreas em que o consumo de energia é crítico para a densidade do rack.

Em termos de performance chegou-se à conclusão que o IBM PpowerPC embarcado de baixa frequecia e de baixo consumo, consistentemente superam os processadores de mais alta frequencia e de alto consumo num fator de 2 a 10.

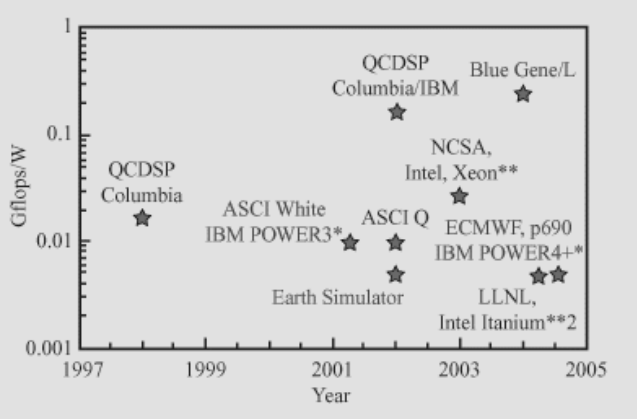


Figura 1.1. Eficiência de consumo.

A figura 1.1 ilustra a eficiência de consumo de alguns supercomputadores recentes. Os dados são baseados no pico total de instruções de ponto flutuante por segundo dividido pelo consumo total de energia,

O sucesso de usar chips de baixo consumo e baixa frequencia só pode ser considerado se o usuário conseguir mais performace escalando para um númeromaior de nós. O ponto-chave são aplicações que tenham um bom comportamento quando escaladas porque sua performace total é aumentada muito mais pelo paralelismo do que por ganhos marginais obtidos por processadores que consomem muito mais e que tenham uma frequancia muito alta.

A importancia de baixo consumo pode ser vista sob vários aspectos. A potencia total de um computador de 360 Tflops baseado num processador convencional de alta performace pode exceder 10 Megawatts e possivelemtne chegar perto dos 20 Megawatts. Como referencia 10 Megawatts é igual ao consumo de energia usado por 11.000 casas nos Estados Unidos. Outro ponto crítico para atingir um numero sem precedentes de escalabilidade é oque se chama de RAS (Reliability, Availability and Serviciability) que foi onde o BG/L teve muito foco.

O BG/L foi desenhado para usar eficientemente a memória distribuida e o modelo de programaçcão message-passing (MPI) juntamento com as linguagens C/C++ e Fortran.

Um simples rack de 90 cm de largura, 90 cm de profundidade e 1.90 m de altura pode conter 1024 dual processors compute nodes consumindo um total de 27.5 Kilowatts de energia o que seria o equivalente a 275 lâmpadas de 100 W.

2.AMBIENTE OPERACIONAL

A arquitetura do BG/L pode ser descrita em alto nível conforme mostrado na Figura 2.1 e consiste de um core computacional contendo 65.536 compute nodes e 1024 I/O nodes.

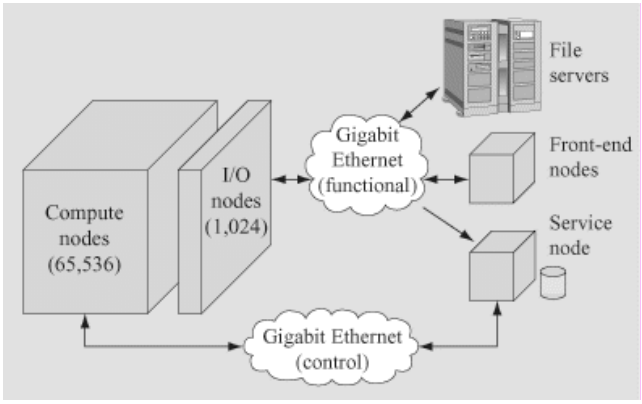


Figura 2.1. Visão em alto nível da arquitetura do BG/L.

2.1Organização do Sistema

Uma rede Ethernet gigabit chamada de *funcional*, conecta o core computacional ao service node (controle da máquina), aos front-end nodes (onde os usuarios compilam, submitem e interagem com seus jobs) e aos servidores de arquivos paralelos.

O service node também está conectado ao core computacional através de outra rede Ethernet gigabit chamada de *control* e que é usada para a manipulação direta do hardware conforme descrito adiante.

O core computacional de 65,536 compute nodes é dividido em 1024 conjuntos lógicos chamados de *psets*. Cada *pset* consiste de um I/O node rodando Linux e de 64 compute nodes rodando um kernel especialmente desenvolvido para o compute node e chamado de CNK. Os *psets* não são entidades fisicas na mas sim lógicas onde são agrupadas logicamente a partir de compute e I/O nodes de uma partição assinalando-se compute nodes à um I/O node em particular. Existe uma certa flexibilidade em assinalar nodes a um pset e essa configuração é parte do setup da máquina. Na prática, configurações altenativas são possíveis com uma distribuição de I/O para compute nodes de 1:8 a 1:128,

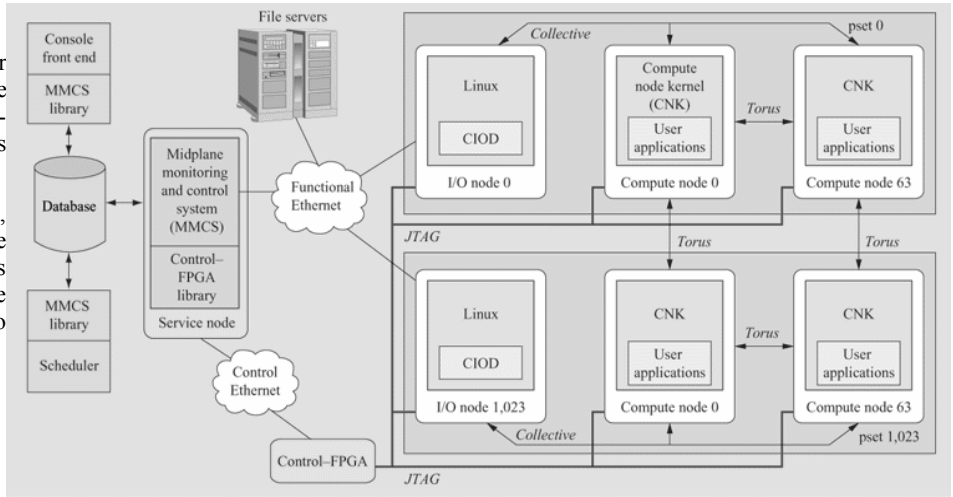


Figura 2.2. Visão em alto nivel de software do Blue Gene/L

ou seja, psets podem ser tão pequenos quanto 8 compute nodes ou grandes como 128 compute nodes condicionados à quantidade de hardware adequado estar presente.

Do ponto de vista de Software, a máquina pode ser dividida em três princiapis componentes:

- Computational Volume: implmentado pelos kernels e as bibliotecas de runtime que suportam a execução das aplicações dos usuários onde os compute nodesde um *pset* são vistos como mecanismos computacionais ligados ao seu I/O node que age como a ponta ou cabeça do *pset*.
- Functional Surface: suporta as compilações das aplicações, inicialização e controloe do programa, debugging e as operações de I/O das aplicações que estão rodando. Os I/O nodes, front-end-nodes e services nodes implmentam a functional surface que é usado para expor a maquina para o mundo externo.
- Control Surface: é implementada exclusivamente no service node que repomsável por operações como boot da máquina, monitoramento de dados do ambiente (temperatura e voltagens) e reportes de erros críticos. Um simples node multiprocessado implementa toda a funcionalidade requerida para um sistema com 65.536 compute nodes.

2.2Software e Sistema no I/O node

Os I/O nodes rodam com Linux e els tem duas missões no BG/L. Eles são reposnsáveis por fazer o controle e a inicialização dos programas nos seus respectivos *psets*. A outra função é que eles também são os responsáveis pelo offload da maioria dos serviços de sistema requeridos pelas aplicações ou jobs que estão nos compute nodes do pset. O código de usuário nunca roda diretamente no I/O node

O Kernel de Linux que roda no I/O node é baseado nas distribuições portadas para o processador IBM PowerPC 440GP (PPC440).

O BG/L usa o core standard do PPC440 e foram necessárias algumas mudanças na sequencia de boot, gerenciamento de interrupções, layout de memória, suporte ao Floatin Point Unit (FPU) e em device drivers do kernel standard. Não existe Basic I/O System (BIOS) nos nodes do BG/L, portanto a configuração do node depois do power-on é parte do Initial Program Load (IPL) que é iniciado pelo service nodeatravés da rede de controle.As rotinas de interrupt e exception-handling foram modificadas para suportar o BG/L Interrupt Controller (BIC) especialmente desenhado para o BG/L. A implementação dos serviços de gerenciamento de memória do Kernel do Linux para os I/O nodes remapeia para o suer space o torus e os devices da collective network. O kernel suporta também o mac da rede Gigabit Ethernet (EMAC4) através de device drivers que foram exptendidos para suportar o save/restore dos resitradores duplos da FPU do BG/L (específicos do BG/L).

Os nodes do BG/L são diskless, então o root file system é carregado como um ramdisk carregado na memória do I/O node durante o IPL. Esse ramdisk contém shells, utilitários, bibliotecas compartilhadas e alguns clientes de rede como *ftp* e *nfs*. Assim que o kernel dá o boot através do ramdisk ele pode usar o client nfs para montar os file systems dos servidores de rede.

As funções carga e inicilização dos programas de usuário, tratamento de sinais, fim do programa, e I/O para arquivos são realizadas utilizando mensagens ponto a ponto entre o compute node e o seu respectivo I/O node através do collective network. Essa funcionalidade é feita pelo Control and I/O Daemon (CIOD) que roda nos I/O nodes. O CIOD faz o controle dos programas e gerenciamento de I/O para todos os compute nodes do pset. Quando inciliza um job, o CIOD carrega o código executável em todos os compute nodes do pset e então envia uma mensagem de start para o CNK desses nodes. Depois que todos os processos estão rodando, o CIOD fica monitornado as mensagens que vem do CNK que contém descrições de operações de I/O a serem feitas pelo CIOD para os processos que estão rodando nos compute nodes. As mensagens entre o CIOD e CNK são sincronas e todas as operações de de arquivo e socket são do tipo blocante.

2.3Software e Sistema no Compute Node

Os compute nodes executam o CNK, um kernel reduzido especialmente desenvovido e que roda como single-user, com duas threads (uma thread por cada processador). O CNK controla o compute node provendo um address space fixo e linear, sem paginação de 512 MB por compute node. O Kernel e o programa do usuário residem no mesmo address spcace sendo que o Kernel reside numa area com memosria protegida inciando no endereço 0, o prgrama de usuário carregado logo em seguida e logo depois o heap e stack. O Kernel protege a sí proprio apropriando-se da MMU (memory management unit). Os recursos físicos, torus e collective network, locks, barriers e scratchpads estão particionados entre o kernel e a aplicação. O torus é totalmente mapeado no user space para conseguir um melhor desempenho na comunicação enquanto um dos dois collective channel fica disponível para o kernel a outra fica para a aplicação.

O CNK implementa os padrões POSIX de interface. A GNU glibc runtime library foi portada para prover um conjunto basico de system calls para file I/O pois essas funções não são feitas diretamente pelo CNK pois são enviadas ao CIOD no I/O node do pset para que sejam executadas e os resultados são recebidos de volta.

O software de message-passing que permite que as aplicações rodando em compute nodes se comuniquem umas com as outras é implementado em 3 níveis:

- O packet layer uma biblioteca que disponibiliza funções para acessar diretamente o hardware de rede do BG/L/
- O message layer que provê o meio de mensagem ponto a ponto com baixa latencia e alta velocidade.
- O message passsing interface (MPI) que é a biblioteca de comunicação disponível a nível de aplicação.

2.4Software e Sistema no front-end node

Os compiladores e debuggers rodam no front-end nodes que na realidade são os unicos nodes onde os usuários conseguem logar. A submissão de jobs é feita também pelo front-end nodes sendo que a carga do job e controle do mesmo é feito pelo service node.

O BG/L traz um ambiente familiar e um conjunto de ferramentas como binutils, glibc, gcc e gdb que foram

portadaspara o BG/L e também foi disoponibilizado um ambiente de cross compiler para suportar como targe o Linux para o I/O node e o CNK para os compute nodes. O laboratório de linguagens da IBM de Toronto no Canadá fez o porte da suite de compiladores XL para trazer o suporte dos compiladores Fortran 90, C e C++ para o BG/L. Uma das caracteristicas principais desses compiladores é a geração de código com suporte automático ao SIMD (Single Instruction Multiple Data)

2.5Software e Sistema no front-end node

A infra-estrutura de controle é um componente crítico no design do BG/L. Os sistemas operacionais que rodam no nodes do BG/L (Linux nos I/O nodes e CNK nos compute nodes) implementam serviços e são responsáveis somente por decisões locais que não afetam a máquina como um todo por isso é necessário um sistema operacional com abrangencia global, que possa tomar decisões coletivas, fazer interface com outros módulos e executatar uma série de serviços de gerenciamento como boot de máquinas, monitoramento do sistema e incilização de jobs. O sistema opercacional que roda no fron-end node chama-se Core Management and Control System (CMCS).

O CMCS contém dois processos principaisonde cada um deles suporta uma diferente área no sistema:

- Midplane Monitoring and Control System (MMCS) usa a restricted control network para manipular o hardware diretamente para configura-lo, incializa-lo e coloca-lo operacional.
- Control and I/O Manager (CIOMAN) usa a Ethernet funcional para comunicar-se com os sistemas operacionais para suportar o job execution.

O monitoramento do sistema é feito com uma combinação de serviços do MMCS e CIOMAN e os serviços dos I/O nodes.

3.PROCESSADOR PowerPC 440

Os chips usados na implementação do BG/L sao dual core PowerPC 440 do tipo SOC (system-on-a-chip) que itegra além dos componentes básicos de um processador, duas unidades independentes de floating point com dupla precisão e os demais componentes de hardware para implementar por exemplo o Torus network e a rede Ethernet.

Cada um dos dois cores foi desenhado para chegar a uma frequencia nominal de 700 MHz. O modelo core do PPC400 pode ser visto na figura 3.1. O processador de 32 bits tem uma implementção superscalar baseado nas especificações do PowerPC. O pipeline de instruçÕes é composto de 7 estágios com duplo fetch de instruções, decode, e out-of-order issue. Ele tem também out-of-order dispatch, execution e completion. O Branch prediction é suportado pelo do Branch History Table (BHT). O Brach target address cache reduz a latencia dos branches. O PPC440 tem 3 pipelines independentes: load/store, inteiro simples e um outro combinado para inteiros complexos, sistema e branch. O sistema tem 32 registradores de 32 bits. O cache de L1 é implementado em separado para instruções e dados cabendo a cada um 32 KB de cache L1 com linhas de 32 bytes e 64-way associativo com round-robin, suportando tanto write-back como writhe-through e não blocante suportando até 4 misses pendentes.. O mmu (memory management unit) é do tipo fully associative com 64 entradas e um TLB (Translation lookaside buffer) que suporta tamanhos variados de páginas e quatro storage atributos que podem ser definidos pelo usuário.

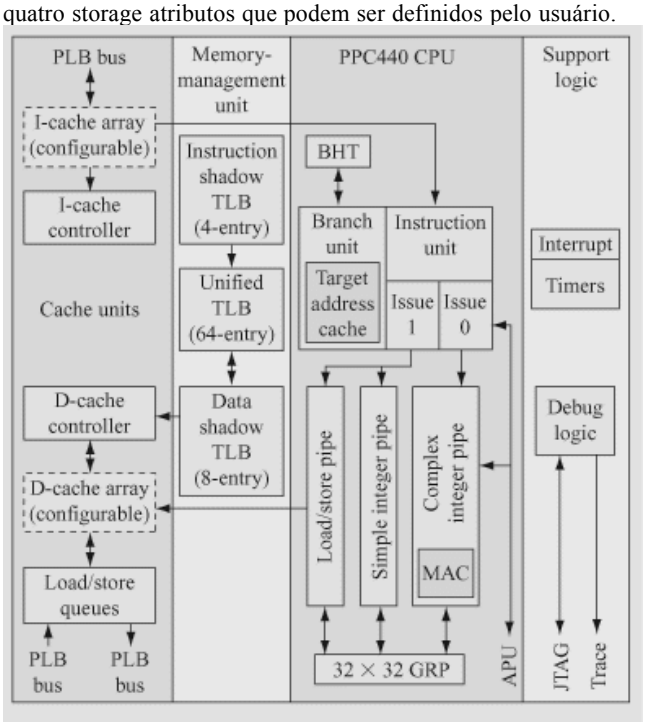


Figura 3.1. Core do processador PowerPC 440

A figura 3.2 mostra os componentes basicos de acesso à memória de um node. O cahce L1 está contido dentro do core do PPC440. O segundo nível (L2R e L2W) são pequenos e basicamente serve como pre-fetch e de buffer de write back para os dados do L1. O cache L3 é grande e dele é esperado uma grande taxa de transferencia com baixa latência de acesso. este cache é usado tanto para dados como para instruções. Finalmente a memoria principal é implementada fora do chip com memoria DDR SDRAM.

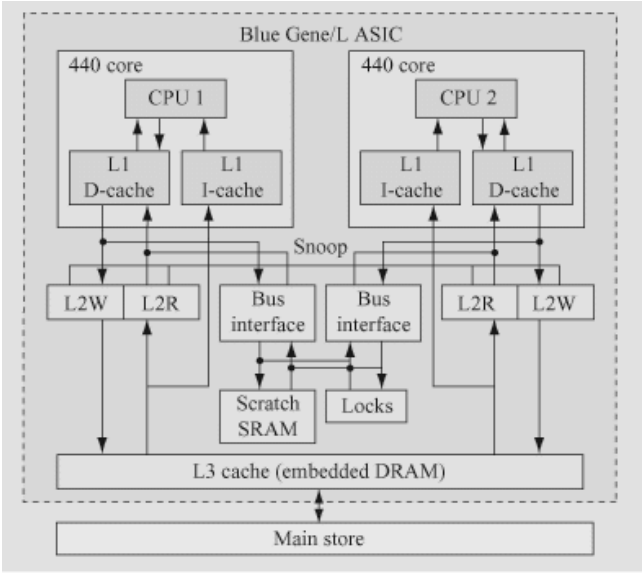
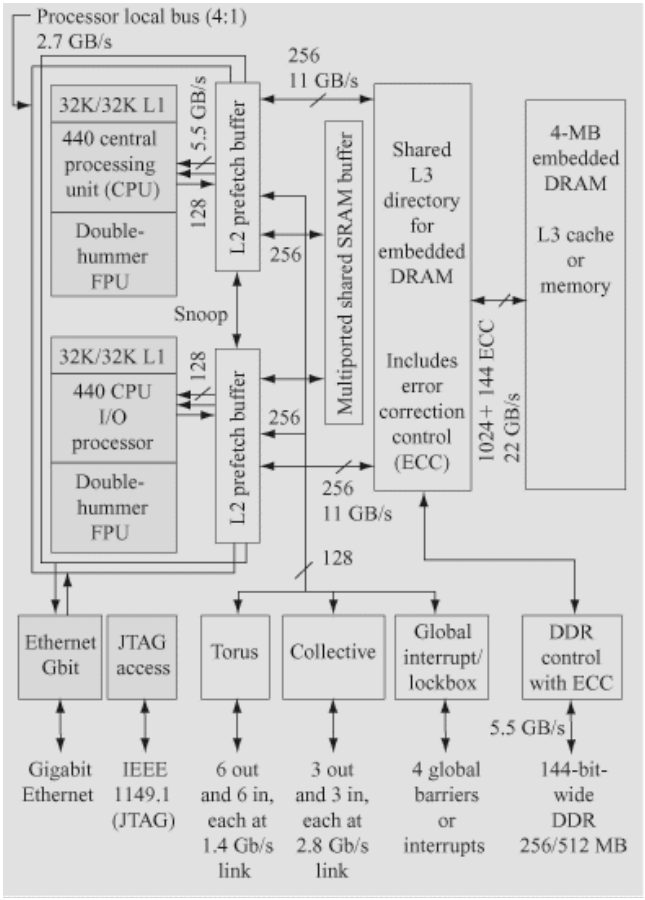


Figura 3.2. Estrutura de cache e memória

A figura 3.3 mostra todos os compoentes implementados dentro do chip PPC440 usado no BG/L. Os boxes com o nome de "Double Hummer FPU" são os módulos de floating point que estão acoplados à cada processador e são capazes de produzir uma performace de pico de operações de floating point por cicle de processador.



Podem ser observados também que todos os componentes de I/O como a Ethernet GB, JTAG, collective network e torus estão implementados dentro do chip o que reduz espaço e consumo.

Um dos recursos mais importantes em processamento paralelo maciço é a rede que interconecta os processadores e que permite que a máquina como uma grande entidade coerente. Essa rede é implementada através da torus network que é uma rede 3D em que cada compute node pode se comunicar com 6 vizinhos na forma 2x2x2 ou xyz. A velocidade de conexão é de 1.4 Gb/s em cada direção. Cada node suporta seis idependentes conexões bidirecionais.

A figura 3.4 mostra as principais inteconexões existentes no BG/L onde: (a) Torus network, (b) Global collective network, (c) BG/L control system network e gigabit network.

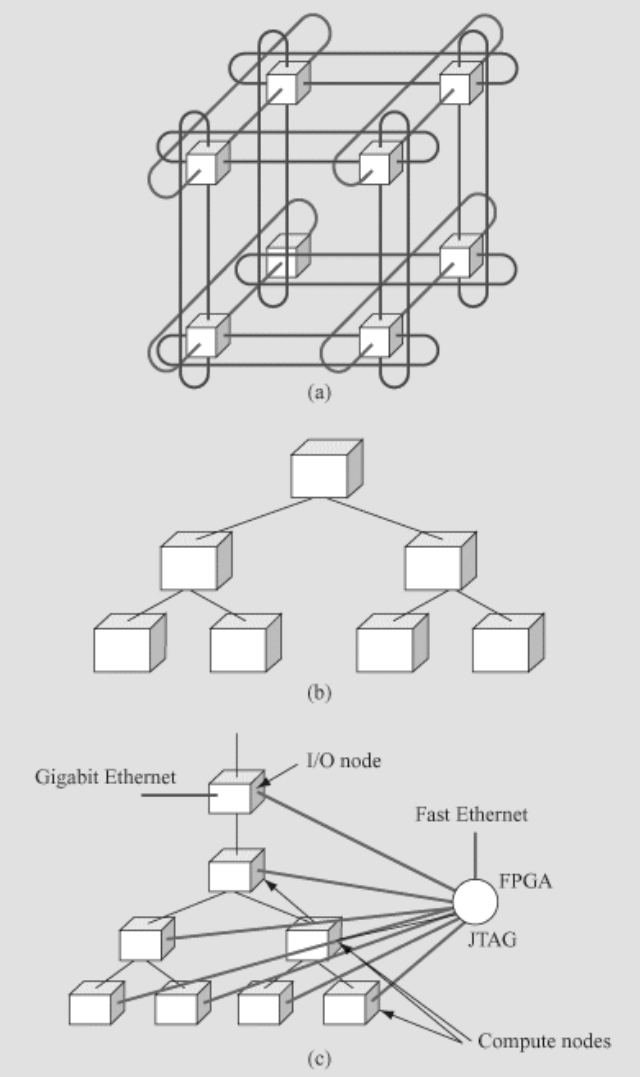


Figura 3.4. Tipos de interconexoes entre nodes do BG/L

4.REFERÊNCIAS

[1] IBM Journal of Research and Development
Vol 49, number 2/3, March/May 2005
http://www.research.ibm.com/journal/rd/492/tocpdf.html

[2] TOP 500 List
http://www.top500.org

[3] LINPACK
http://www.netlib.org/linpack

[4] POWER ORG
www.power.org

[5] IBM Redbooks
http://www.redbooks.ibm.com/cgi-bin/searchsite.cgi?query=blue+AND+gene