IBM

Modelagem de Proteínas com Blue Gene/L

Avanços Científicos do Mundo Real através de Modelagem e Visualização de Dados em um Supercomputador

O supercomputador Blue Gene®/L fornece aos cientistas uma potência computacional de ponta e ferramentas de visualização de dados complexos das quais eles precisam para estarem sempre à frente em suas disciplinas. Saiba como esta tecnologia permite que os especialistas em biologia molecular computacional criem simulações de dobramento e de desdobramento de proteínas para entenderem melhor essas moléculas complexas.

Chris Ward se juntou ao IBM UK Development Laboratories em Hursley, Inglaterra, em 1982 com um diploma de engenharia da Cambridge University. Ele trabalhou no desenvolvimento de vários produtos da IBM, de arquivos de disco a middleware de marca. Ele é um privilegiado por trabalhar com uma tecnologia que é tão valiosa para os futuros clientes da IBM quanto o IBM WebSphere Software e o IBM Lotus Software são para atuais clientes da IBM.

Ruhong Zhou é Cientista da Equipe de Pesquisas do Computational Biology Center/IBM Thomas J. Watson Research Center e Professor Adjunto no departamento de química da Columbia University. Recebeu seu Ph.D. do Bruce Berne em química da Columbia University em 1997. Se juntou ao IBM Research em 2000 depois de passar dois anos e meio trabalhando com Richard Friesner (Columbia) e William Jorgensen (Yale) em campos de força polarizáveis e mecanismos de ligação de proteínas ligantes. É autor de 80 publicações journal e 7 patentes, participou como convidado de inúmeras palestras nas principais conferências e universidades, além de presidir inúmeras conferências sobre química e biologia computacionais e biofísica. Ganhou o Hammett Award em 1997 da Columbia, DEC Award em 1995 da American Chemical Society on Computational Chemistry e o Outstanding Technical Achievement Award em 2005 e 2008 da IBM. Seus atuais interesses em pesquisas incluem o desenvolvimento de algorítmos e métodos novos para biologia computacional e bioinformática, além de simulações em grande escala para dobramento de proteínas, ligação de receptor ligante e previsão de estrutura protéica.

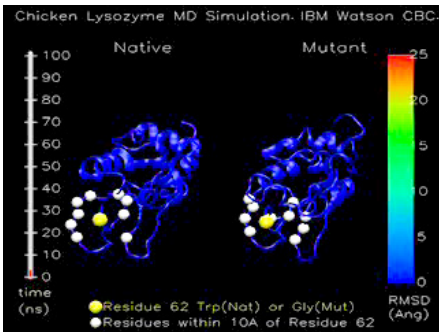
09/Jun/2009

Em 2001, cientistas pesquisadores da IBM iniciaram o design de uma nova família de servidores, comercializados hoje como o IBM System Blue Gene®. Esses servidores estão disponíveis desde 2004—primeiro o Blue Gene/L (sobre o qual falamos neste artigo), e o Blue Gene®/P.

A família Blue Gene de supercomputadores foi projetada para oferecer desempenho em ultra-escala com um ambiente de programação padrão; ela também foi projetada para ter desempenho eficiente no consumo de energia, resfriamento e área ocupada. Muitas universidades, governos e laboratórios de pesquisa comerciais utilizam o Blue Gene para estudos de computação em radioastronomia, dobramento de proteínas, pesquisas climáticas, cosmologia e desenvolvimento de medicamentos. O sistema está fazendo uma mudança, por ordem de magnitude, na maneira como a ciência pode ser realizada, pois oferece uma ferramenta com custo reduzido para o design e a execução de versões alternativas de modelos complexos.

Neste artigo, apresentamos alguns dos progressos que foram feitos por um dos projetos ligados à modelagem de proteínas. A Figura 1 mostra a escala de trabalho que podemos realizar agora, graças ao poder do Blue Gene/L. . A configuração inicial começa na estrutura cristalina da lisozima (consulte Recursos para obter a fonte).

Figura 1. Parte do Total de Dez Microssegundos de Vida dentro de uma Célula Viva (assista ao vídeo)



Proteômica: A Economia da Proteína

Proteínas são macromoléculas biológicas que são um componente essencial dos organismos e que participam de cada processo dentro das células. Muitas proteínas são enzimas que catalisam reações bioquímicas; algumas estão envolvidas em sinalização celular e resposta imune; muitas outras possuem funções estruturais e mecânicas para músculos e citoesqueletos. Dois exemplos ilustram como as proteínas são difusas e importantes:

- Uma proteína é responsável pela "vermelhidão" do sangue; ela transporta oxigênio dos pulmões para todas as outras partes do corpo.
- Outra proteína é responsável pela resposta do corpo humano ao veneno de um sumagre venenoso; extremamente irritante, mas normalmente não é prejudicial.

Existem centenas de milhares de proteínas envolvidas com a vida na Terra. A proteômica é o estudo de como as proteínas trabalham, como elas interagem entre si e como sua diversidade e especialização evoluem entre os organismos vivos ao nosso redor. Neste artigo, vamos descobrir o que são as proteínas, como elas são feitas e como elas afetam os sistemas que habitam.

O DNA é um componente de armazenamento de informações em cada célula de cada planta e animal. Ele armazena informações como uma sequência de blocos de construção químicos (nucleotídios) que chamamos de **A**, **C**, **T**e **G** (para adenina, citosina, timina e guanina no DNA e uracila que substitui a timina no RNA). A uma certa distância, esses blocos de construção parecem bastante semelhantes, portanto, cada parte de um DNA que você vê tem a mesma forma geral—a famosa Dupla Hélice de Watson e Crick.

Para ler as informações no DNA, o DNA se desenrola e outra molécula chamada RNA é formada pela apresentação do padrão interno. Em vez pressionar uma chave na almecega, agora você tem a imagem de uma chave na almecega. Essa molécula de RNA é apresentada em seguida como um projeto ao ribossomo, uma proteína que se comporta como uma fábrica multifuncional. O ribossomo lê o código A/C/T/G em grupos de três, o que nos permite criar um "alfabeto" de 64 letras.

Vinte dessas "letras" correspondem a aminoácidos, os blocos de construção das proteínas. Esses aminoácidos são provenientes principalmente daquilo que comemos (os humanos não podem produzir todos os aminoácidos de que precisamos e, portanto, devem obter os outros, chamados de aminoácidos "essenciais", da comida). Cada aminoácido possui "cabeça" e "cauda". O ribossomo encontra o aminoácido apropriado para cada "letra" e os monta da cabeça à cauda em sequência; outras "letras" indicam quando começar e quando parar. A sequência linear resultante dos aminoácidos é uma

molécula de proteína recém-inventada, formada precisamente de acordo com o código impresso na seção do DNA que foi utilizado.

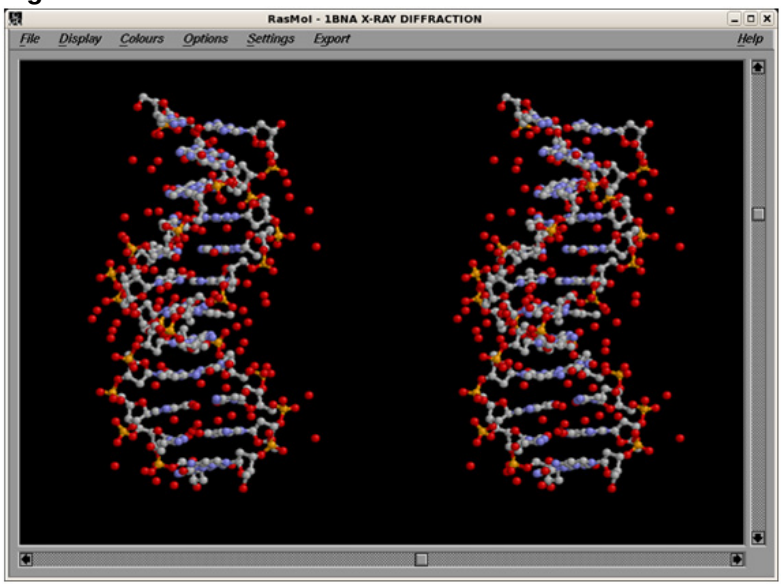
A tensão e o esforço entre os átomos na molécula de proteína, as interações com água levemente salgada na célula e as vibrações aleatórias que você chamaria de *calor* fariam a molécula de proteína se "dobrar" adquirindo uma forma característica.

As moléculas de proteína são totalmente estáveis; algumas delas podem permanecer inalteradas por centenas de anos e suportar temperaturas de centenas de graus, o que mataria o organismo que as compôs. Elas se mantêm em um estado bruto até serem desnaturadas por produtos químicos fortes, alta pressão, calor ou frio ou até se tornarem alimento para outros seres vivos.

A forma e a maneira como ela varia com o tempo, temperatura e moléculas adjacentes determinam o que a molécula de proteína vai fazer—se ela vai transportar oxigênio, se vai fazê-lo ter uma reação alérgica a um sumagre venenoso ou se vai fazer qualquer outra que possa acontecer em uma escala minúscula.

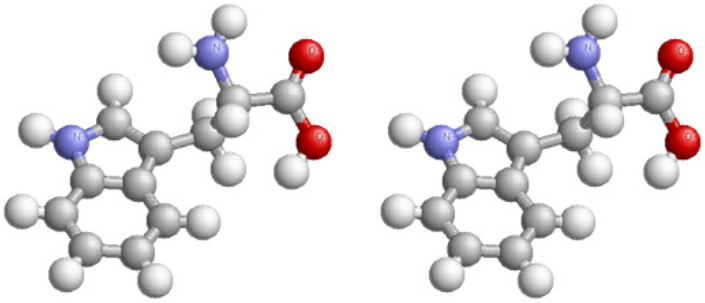
A Figura 2 demonstra o modelo familiar de bolas e varetas de DNA (a imagem é um par estéreo; consulte Recursos para obter a fonte da imagem):

Figura 2. O Modelo de Bolas e Varetas de DNA



A Figura 3 mostra o triptofano, um dos 20 aminoácidos padrão (a imagem é um par estéreo; consulte Recursos para obter a fonte da imagem).

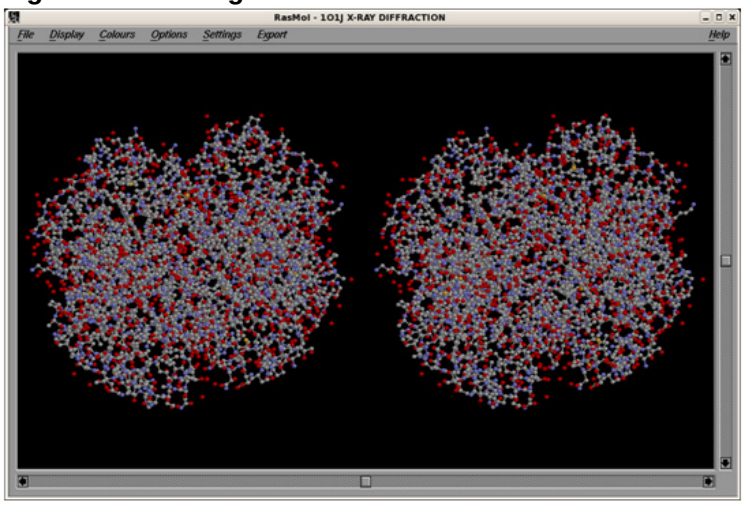
Figura 3. Triptofano, um dos 20 Aminoácidos Padrão



Os aminoácidos são acumulados nas proteínas através da desconexão do grupo O-H (lado direito da Figura 3) de uma molécula, separando-se o H do N (parte superior da Figura 3) de outra molécula e juntando as moléculas. O grupo H-O-H mantido é uma molécula de água. Todos os aminoácidos têm esse agrupamento atômico característico (parte superior direita da Figura 3).

A Figura 4 fornece um aspecto visual da hemoglobina de proteína (a imagem é um par estéreo; consulte Recursos para obter a fonte da imagem).

Figura 4. A Hemoglobina de Proteína



A hemoglobina é um total de 574 moléculas de aminoácido em 4 subunidades. A hemoglobina, com seus átomos de ferro associados (a forma como eles são acumulados na proteína estão além do escopo deste artigo), transporta oxigênio para toda a circulação sanguínea. Um sistema de transporte de oxigênio é possível com apenas os átomos de ferro, mas ele é muito mais eficaz com a "gaiola" de proteínas fornecida pela estrutura da hemoglobina. Se colocar essa imagem em um visualizador estéreo, você conseguirá uma estrutura atômica em 3D; para algo mais complexo, precisamos de uma maneira diferente de visualizar o que está acontecendo.

Motivações Comerciais e Acadêmicas

Cada vez mais, os avanços no design farmacêutico e na proteção da saúde pública estão vindo de um entendimento melhor dos blocos de construção básicos de vida, como as proteínas. Um

O que É um Tipo Selvagem?
Um tipo selvagem é a forma típica que um organismo, gene, classe ou característica tem por natureza. Se

tópico atual é *agregação e mal dobramento de proteína*—se uma proteína for dobrada de uma forma diferente da pretendida, o resultado costuma produzir proteínas inativas com diferentes propriedades, o que pode levar a doenças neurodegenerativas, como Mal de Alzheimer, Doença de Creutzfeldt-Jakob, Encefalopatia Espongiforme Bovina (Vaca Louca), Mal de Huntington e Mal de Parkinson, fibrose cística e outras amiloidoses.

Entender o que faz as moléculas de proteínas mudarem sua forma de dobramento útil para uma forma diferente é um tópico ativo nas pesquisas de tratamentos para essas doenças significativas. Experiências recentes lideradas por Chris Dobson e seus colaboradores na Cambridge University (consulte Recursos para obter um link) mostraram que fibrilas e amilóides podem ser formadas não apenas a partir de peptídes beta-amilóides tradicionais, mas também a partir de quaisquer proteínas (como lisozimas) dadas as condições apropriadas. De fato, uma única mutação (W62A) na proteína lisozima pode deixar a proteína em um estado muito menos estável em comparação com o tipo selvagem (consulte a barra lateral); ela também causar seu mal dobramento e formar possíveis amilóides na solução de uréia devido à perda de "interações hidrofóbicas de longo alcance" importantes.

estiver se referindo ao *fenótipo* (as características observáveis de um organismo, geralmente a expressão de genes e fatores ambientais), o tipo selvagem caracteriza os traços mais comuns na população natural. Se estiver se referindo ao *genótipo* (a composição genética não observável), ele define o alelo em cada lugar necessário para a produção do fenótipo do tipo selvagem. Os tipos selvagens não são nem dominantes e nem recessivos. Uma bom antônimo para tipo selvagem é *mutante*.

Cientistas ainda não sabem como esse único resíduo de W62 pode desempenhar um papel tão importante nas interações hidrofóbicas de longo alcance durante o processo de dobramento e depois se deslocar para a superfície presumivelmente a partir de um local de nucleação por razões de funcionamento. Isso oferece uma oportunidade exclusiva para um melhor entendimento dos efeitos dessa mutação única, bem como o mecanismo por trás das doenças supramencionadas relacionadas à agregação e ao desdobramento de proteínas.

A tecnologia Blue Gene/L pode ser utilizada para a abordagem desses tipos de doenças, pois ele fornece uma maneira com custo reduzido (e mais rápida) de se modelar os efeitos do dobramento e do desdobramento de proteínas.

Então, o que Estamos Modelando?

O vídeo do qual a Figura 1 foi capturada é uma visualização de parte de uma sequência de um desdobramento de uma proteína lisozima devido a uma única mutação. A lisozima é uma proteína que faz parte do sistema imunológico humano; quando está funcionando corretamente, ela perfura as paredes da célula de uma bactéria invasora e a destrói.

Uma mutação única, uma sequência diferente no DNA, faz o ribossomo utilizar um aminoácido diferente ao construir a molécula de lisozima. A teoria é que esse aminoácido diferente afeta a forma como a lisozima se dobra e que essa molécula de lisozima com forma diferente age de forma diferente na perfuração das paredes das células bacterianas. Ao entendermos essa mudança, podemos desenvolver medicamentos ou outras formas de terapia que ajudarão os indivíduos com essa mutação a se recuperar de doenças bacterianas.

Como parte do trabalho, armazenamos as posições e velocidades de cada átomo em uma molécula de

lisozima, bem como aqueles com aproximadamente 10.000 moléculas de água e uréia (essa simulação é feita em uma solução de uréia de molar 8 para imitar experiências) na memória do computador. Existem várias maneiras de se modelar as forças entre átomos; utilizamos uma variante de um modelo de *bolas e varetas* para forças ligadas, com um modelo de lei do quadrado inverso para forças eletrostáticas entre átomos carregados e um modelo de atração/repulsão para átomos que estão próximos uns dos outros, mas sem ligação covalente. O modelo é executado como uma série temporal. Em cada etapa temporal, calculamos as forças em cada átomo e, então, atualizamos as velocidades e as posições de acordo com a Segunda Lei de Newton.

Em cada etapa temporal (muito pequenas, com cerca de 1 femtossegundo), existem em princípio centenas de milhares de forças para serem calculadas. Também queremos poder executar simulações longas o suficiente (microssegundos) para modelar movimentos interessantes—é claro, isso significa que essa abordagem se tornou prática apenas recentemente, mesmo com os maiores computadores que nós sabemos como construir. Para obter mais detalhes sobre o que fazer e algumas abordagens alternativas, consulte o link para "Destruction of long-range interactions by a single mutation in lysozyme" em Recursos.

Equipando o Laboratório

No IBM Watson Research Lab em Yorktown, Nova Iorque, temos 20 racks de servidores BlueGene/L. Cada rack contém 1.024 chips microprocessadores dual-core PowerPC®; cada microprocessador está conectado a 512 MB de RAM. Para cada 64 chips nesta *rede de computadores*, existe um microprocessador adicional conectado a um link de Ethernet de 1 Gbps. Esses 320 links de Ethernet são conectados através de comutadores Ethernet padrão a máquinas IBM Power Systems padrão com discos, fitas, compiladores de linguagens e software de controle de tarefa.

Esse trabalho de modelagem de lisozima utilizou uma média de quatro racks de processadores BlueGene/L por vários meses para gerar um agregado de mais de 10 microssegundos de dados dinâmicos moleculares. Periodicamente, o aplicativo grava as posições e as velocidades de todos os átomos sob simulação (parte deste fluxo de informações foi utilizada para produzir o vídeo sintético mencionado antes). Sempre que for necessário reiniciar a execução de simulação, um conjunto apropriado de posições e velocidades pode ser recarregado. Pode ser necessário reiniciar após um encerramento planejado, após uma falha da máquina não planejada ou para reproduzir um evento modelo de interesse científico com uma granularidade de etapa temporal diferente.

Executando o Modelo

O aplicativo é inicializado nos nós do Blue Gene/L por um mecanismo semelhante ao envio de tarefa MPICH (MPICH é uma implementação móvel disponível gratuitamente de MPI, a message-passing interface; consulte Recursos para obter um link). Cada processador no cluster fornece um ambiente de sistema de arquivos POSIX para o aplicativo. Dados podem ser configurados em um sistema de arquivos IBM General Parallel File System (GPFS) para o aplicativo ler; quando o aplicativo gravar os resultados, esses resultados também deverão ir para lá para uso externo.

Para aplicativos de modelagem de série temporal como esta, é normal ler as condições iniciais a partir do sistema de arquivos e gravar "capturas instantâneas" periódicas do estado do modelo para o sistema

de arquivos.

O que Isso tudo nos Oferece?

O vídeo é uma visão rápida de um mundo que nunca foi visível antes. É claro, nós não sabemos se isso representa a verdade—os cientistas sempre têm que comparar o que um modelo mostra com o que eles vêem no mundo real. Ver como a lisozima se dobra mal na realidade ainda é um sonho; mesmo "vendo" parte das conformidades "corrigidas" significa preparar amostras e colocá-las sob um microscópio eletrônico ou, possivelmente, causar a cristalização de inúmeras moléculas de lisozima e depois utilizar espectroscopia de difração de raio X. Porém, essas técnicas experimentais normalmente não nos dão um insight de como a proteína pode mudar.

Portanto, as atuais simulações em grande escala nos oferecem uma janela exclusiva para olharmos os detalhes dos movimentos moleculares e as mudanças críticas envolvidas nos desdobramentos relacionados a doenças. Felizmente, a disponibilidade da tecnologia que pode fazer isso acontecer irá desafiar todos os limites e fará avanços na modernidade dos estudos de amiloidoses. Isso também pode ser usado para treinar a próxima geração de cientistas para resolver esses tipos de problemas utilizando essa nova forma como método principal para esse tipo de pesquisa.

Prevendo o Futuro

De fato, não seríamos tão corajosos a ponto de tentar adivinhar o amanhã, mas nos arriscaríamos a pensar que a computação Blue Gene continuará seguindo um caminho para o desenvolvimento (utilizamos a versão L; o Blue Gene/P disponível foi atualizado para 4 processadores por chip, Ethernet de 10 Gbps e vários outros aprimoramentos). O custo para deixar a aritmética mais computacional e o custo de um armazenamento mais rápido (ambos fortemente associados às tarefas de visualização de dados descritas neste artigo) provavelmente continuarão caindo—e devem, pois existem vários mundos dignos da modelagem avançada que os cientistas precisam fazer, tanto para pesquisas públicas quanto para os negócios para trazer produtos para o mercado.

O modelo de lisozima descrito aqui investiga apenas a molécula do novo campo da biologia computacional. Existem mais de 50.000 proteínas cujas estruturas são catalogadas no Protein Data Base público; existem milhões de possíveis componentes úteis na área farmacêutica para serem analisados; e existem várias doenças humanas conhecidas como sendo relacionadas às proteínas e aos seus defeitos. E nós não estamos nem considerando a variedade de outras áreas de pesquisas que podem ser beneficiadas pela modelagem nesta escala. O trabalho do Blue Gene está apenas começando.

Recursos

Aprender

Um apanhado geral das pesquisas relacionadas ao Blue Gene/L pode ser encontrado na página do projeto IBM Blue Gene. Outros recursos e componentes da solução Blue Gene incluem:

A página da IBM Solução Blue Gene/P

O IBM General Parallel File System

Tudo que você poderia saber sobre o compilador IBM XL C/C++ Advanced Edition para Blue Gene

IBM Redbooks sobre tecnologias Blue Gene

E uma ilustração do Blue Gene utilizada por um dos autores

O RCSB Protein Data Bank (PDB) é um archive para o estudo de macromoléculas biológicas com informações sobre conjuntos complexos, ácidos nucleicos e estruturas de proteínas experimentalmente determinados. Recursos educacionais incluem coisas interessantes, como A Molécula do Mês.

Os dados de origem para a Figura 1 são do PDB, The 1.33 A structure of tetragonal hen egg white lysozome.

Os dados de origem para a A Figura 2 são do PDB, Structure of a B-DNA dodecamer: conformation and dynamics.

Figura 3 é uma cortesia da biblioteca MathMol hospedada na New York University.

Os dados de origem para a Figura 4 são do PDB, Deoxy hemoglobin (A-GLY-C:V1M,L29F,H58Q; B,D:V1M,L106W).

O grupo de Chris Dobson posta links para mais pesquisas em biologia molecular.

"Destruction of long-range interactions by a single mutation in lysozyme" (R. Zhou, M. Eleftheriou, A. Royyuru, B. J. Berne; Proc. Natl. Acad. Sci., 2007) fornece mais informações sobre a abordagem de modelagem utilizada nessas simulações.

"Parallel implementation of the replica exchange molecular dynamics algorithm on Blue Gene/L" (M. Eleftheriou, A. Rayshubski, J. W. Pitera, B. G. Fitch, R. Zhou, R. S. Germain; IEEE, 2006) explica algumas das técnicas matemáticas utilizadas para a simulação.

MPICH2 é o próximo estágio da MPICH, a implementação totalmente móvel de alto desempenho (e gratuita) do padrão Message Passing Interface (MPI).

O Argonne Leadership Computing Facility possui um programa colaborativo que oferece o tempo do Blue Gene/P para a comunidade de ciência da computação.

O aplicativo de modelagem está disponível para demonstração em IBM Innovation Centers em todo o mundo.

Na zona Linux do developerWorks, descubra mais recursos para desenvolvedores Linux e examine nossos artigos e tutorias mais populares.

Consulte todas as dicas de Linux e tutoriais de Linux no developerWorks.

Mantenha-se atualizado com eventos técnicos e Webcasts do developerWorks.

Obter produtos e tecnologias

Open Discovery é uma distribuição Live Linux baseada em Fedora Core de

ferramentas de software de bioinformática, licenciada sob a Academic Free License (AFL), que pode lidar com tudo, desde análise de sequência até tarefas dinâmicas moleculares. Ele pode ser inicializado a partir do DVD ou da chave de armazenamento USB e apresenta persistência de dados. Muito obrigado ao Department of Bioinformatics, SRM University, campus Ramapuram, Chennai, Índia.

Algumas das ferramentas integradas aos aplicativos descritos neste artigo incluem 3D Fast Fourier Transform Library for Blue Gene/L e Custom Math Functions for High Performance Computing de Chris Ward.

Com o software de teste da IBM, disponível para download diretamente a partir do developerWorks, construa seu próximo projeto de desenvolvimento no Linux.

Discutir

Envolva-se na Comunidade do My developerWorks; com seu perfil pessoal e sua página inicial customizada, você pode padronizar o developerWorks de acordo com seus interesses e interagir com outros usuários do developerWorks.