Arquitetura de Alto desempenho

Introdução à computação de alto desempenho

António Rui Borges

Resumo

- *Computação de alto desempenho*
- *Noções básicas de arquitetura de uma máquina paralela*
- *Decomposição paralela*
- *Ferramentas a serem usadas para codificar aplicações paralelas*
- *Leitura sugerida*

Computação de alto desempenho - 1

A área de *computação de alto desempenho* (HPC) está sempre mudando à medida que novas tecnologias e processos se estabelecem. Em geral, refere-se ao uso de vários processadores ou clusters de computadores fortemente acoplados para executar simultaneamente tarefas de uso intensivo de computação com alto rendimento e eficiência. É comum incluir no conceito de HPC não apenas a arquitetura do computador, mas também um conjunto de elementos como sistemas de hardware, ferramentas de software, plataformas de programação e paradigmas de programação paralela.

Ao longo da última década, o HPC evoluiu significativamente, nomeadamente devido ao surgimento de arquiteturas heterogéneas CPU-GPU, o que levou a uma mudança fundamental de paradigma na programação paralela.

Computação de alto desempenho - 2

Principais sites de supercomputadores (em novembro de 2021)

Adaptado de: https://www.top500.org/lists/2021/11/

Sistema nome	Características do nó / Interconectar	Localização	Número de nós/ núcleos Memória total	Alta performance Marca Linpack PF-flops	Poder (MW)	Operativo Sistema
Supercomputador Fugaku Fujitsu A64FX 48C	A64FX 48C 2,2 GHz Tofu Interconexão D	Centro RIKEN de Ciência Computacional Kobe, Japão https:// www.r-ccs.riken.jp/en/	158 976/7 630 848 48 + 2 núcleos/nó 5,09PB	442,01	29,90	Chapéu vermelho Linux
Cume Sistema de energia IBM AC922	IBM POWER9 22C 3,07 GHz NVIDIA Volta GV100 Mellanox EDR Infiniband de trilho duplo	Laboratório Nacional de Oak Ridge Oak Ridge, Estados Unidos https://www.ornl.gov	4 608/2 397 824 2 (22 núcleos) + 6 GPU/nó 2,80PB	143,50	9,78	Chapéu vermelho Linux
Serra Sistema de energia IBM S922LC	IBM POWER9 22C 3,07 GHz NVIDIA Volta GV100 Mellanox EDR Infiniband de trilho duplo	Laboratório Nacional Lawrence Livermore Livermore, Estados Unidos http://www.llnl.gov	3.022/1.572.480 2 (22 núcleos) + 4 GPU/nó 1,38PB	94,64	7,44	Chapéu vermelho Linux
Sunway TaihuLuz (Poder Divino, a luz do Lago Taihu) MPP Sunway	Sunway SW26010 260C 1,45 GHz Sunway	Centro Nacional de Supercomputação Wuxi, China http://www.nsccwx.cn	40 960/10 649 600 4 (1+64 núcleos) / nó 1,31PB	93,01	15,37	Proprietário Baseado em Linux
Perlmutter HPE Cray EX235n	AMD EPYC 7763 64C 2,45 GHz NVIDIA A100 SXM4 Estilingue-10	Laboratório Nacional Lawrence Berkeley Berkeley, Estados Unidos https://www.nersc.gov/systems/perlmutter/	4500/761 856 1/2 núcleo + 4 GPU/nó 0,42PB	70,87	SO HPE Cray 2,59	

Computação de alto desempenho - 3

Supercomputador Fugaku
De: https://www.r-ccs.riken.jp/en/

Computação de alto desempenho - 4

Fujitsu A64FX 48C De:
https://www.r-ccs.riken.jp/en/fugaku/about/

Computação de alto desempenho - 5

Prefixos SI para ordens de grandeza em unidades

Prefixo		***Ordem de magnitude***		***Desempenho de computação* (FLOPS)**	***Tamanho da memória* (B)**	
K (quilo)	Ki	103	2^{10}	KFLOPS	KB	KiB
M (mega)	Mi	106	2^{20}	MFLOPS	MBMiB	
G (giga)	Gi	109	2^{30}	LOPS	GB	GiB
T (tera)	Ti	1012	2^{40}	FLOPS	tb	TiB
P (peta)	Pi	1015	2^{50}	LOPS	OP	PiB
E (exa)	Eo	1018	2^{60}	EFLOPS	BE	EoB
Z (zeta)	Zo	1021	2^{70}	ZFLOPS	ZB	ZoB
Y (yota)	Sim	1024	2^{80}	YFLOPS	YB YiB	

A comunidade de supercomputação pretendia alcançar EFlops até 2020 e está com o objetivo de alcançar ZFLOPS até 2030.

Principais áreas de aplicação de supercomputação

- cosmologia, astrofísica e astronomia
- química computacional, biologia e engenharia
- ciência da computação
- ciências da terra e materiais
- previsão do tempo
- ciência e tecnologia da informação geográfica
- segurança global
- fusão nuclear
- armas e integração complexa

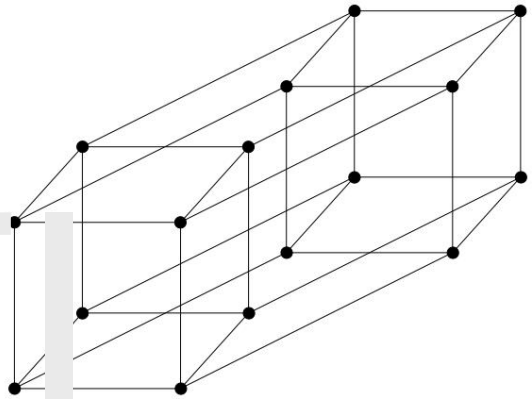
Noções básicas de arquitetura de uma máquina paralela - 1

Os computadores atuais de alto desempenho estão nas máquinas paralelas de memória distribuída de nível superior. Eles podem ser considerados como vastos clusters de nós de processamento (PNs) interconectados por alguma topologia de rede. A razão para isso é *a escalabilidade,* a capacidade de o desempenho do sistema aumentar à medida que novos nós são anexados a ele.

Noções básicas de arquitetura de uma máquina paralela - 2

Topologias de interconexão comuns

malha de toro



hipercubo

árvore gorda

Noções básicas de arquitetura de uma máquina paralela - 3

As principais preocupações sobre a topologia de interconexão são duplas

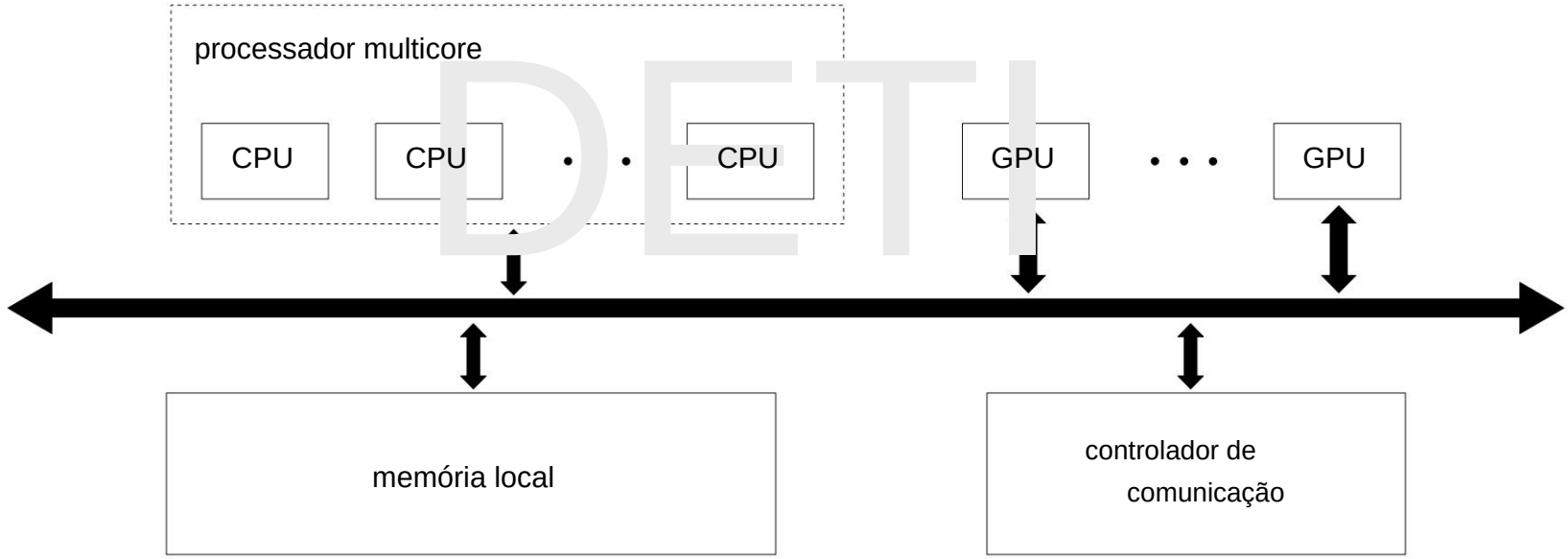
- manter o número de conexões por nó pequeno à medida que o número de nós de processamento no cluster aumenta
- manter o tempo de comunicação e a largura de banda constantes à medida que o número de nós de processamento no cluster aumenta.

Tanto na malha toróide quanto no hipercubo, todas as conexões são ponto a ponto e, como tal, possuem largura de banda fixa. O número de conexões por nó é sempre quatro no primeiro caso e log2 *n* no último caso, onde *n = 2k* é o número de nós de processamento no cluster. O tempo de comunicação, entretanto, depende da localização dos nós de comunicação, sendo no máximo *ÿn* e log2 n, respectivamente, do tempo de comunicação equivalente entre dois nós adjacentes.

Uma árvore gorda, por outro lado, é uma rede hierárquica que tenta manter a mesma largura de banda em todas as bisseções. Todos os nós de processamento transmitem na velocidade da linha se os pacotes estiverem distribuídos uniformemente ao longo dos caminhos disponíveis. Desde uma única conexão por nó é necessário e, usando switches de porta k, $k^3/4$ nós de processamento podem ser anexado a ele, apresenta boas propriedades de escalabilidade.

Noções básicas de arquitetura de uma máquina paralela - 4

Nó de processamento



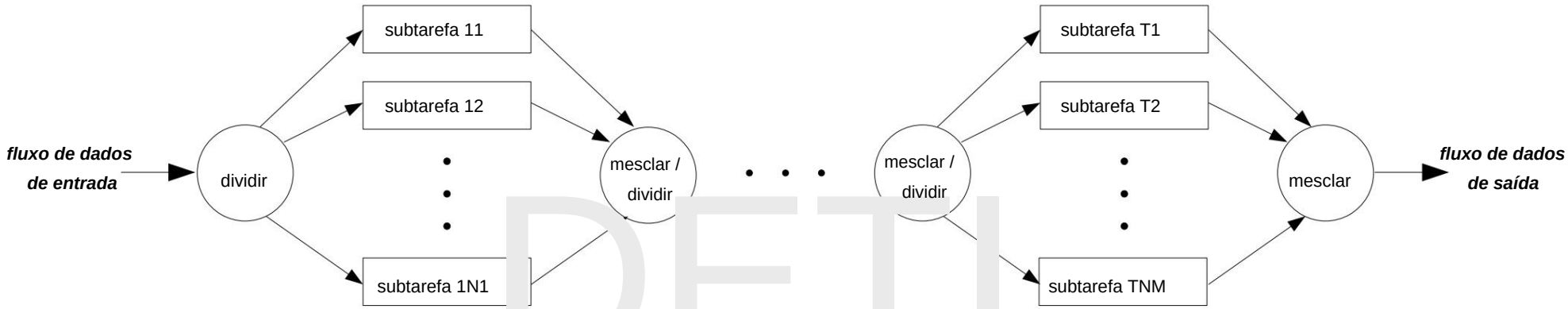
Noções básicas de arquitetura de uma máquina paralela - 5

Um nó de processamento típico consiste em um ou dois soquetes de CPU multicore e duas ou mais GPUs de muitos núcleos que deram origem ao nome *computação heterogênea.* quando se refere a esse tipo de arranjo.

O código da CPU é responsável, neste contexto, por gerenciar o ambiente, o código e os dados da GPU, antes de carregar as tarefas de computação intensiva no dispositivo. A computação GPU não se destina a substituir a computação CPU. As CPUs são otimizadas para cargas de trabalho dinâmicas, marcadas por sequências curtas de operações computacionais e controle de fluxo imprevisível. Por outro lado, as GPUs visam o outro lado do espectro: cargas de trabalho dominadas por tarefas computacionais com controle de fluxo simples.

Assim, as arquiteturas de computação paralela heterogênea CPU+GPU evoluíram porque a CPU e a GPU possuem atributos complementares que permitem que os aplicativos tenham melhor desempenho usando ambos os tipos de processadores.

Decomposição paralela - 1



Normalmente, a decomposição paralela é orientada por dados.

Pedaços de dados do fluxo de entrada são alimentados em um pipeline de operações do estágio T. Em cada estágio, os dados são divididos ainda mais para que as operações possam ser realizadas de forma independente em partes mútuas e exclusivas do bloco que está sendo processado. Entre os estágios, os blocos de dados podem sofrer reorganização.

Decomposição paralela - 2

Algoritmos paralelos podem ser projetados com vários graus de granularidade. *Granularidade* pode ser definido como a forma como as operações paralelas são expressas. Nesse sentido, não é possível expressá-los sem pensar na plataforma de hardware onde o código será executado.

O paralelismo é, assim, organizado em três categorias principais

- *paralelismo refinado* – as operações paralelas são expressas no nível variável, assume que uma instrução é executada simultaneamente em vários conjuntos de dados, um A arquitetura SIMD (instrução única – dados múltiplos) é considerada
- *paralelismo de granulação média* – operações paralelas são expressas no nível do thread dentro de um processo, uma arquitetura MIMD (múltiplas instruções – múltiplos dados) do tipo memória compartilhada é considerada
- *paralelismo de granulação grossa* – as operações paralelas são expressas no nível do processo, sendo considerada uma arquitetura MIMD (múltiplas instruções – múltiplos dados) do tipo memória distribuída.

Na computação de alto desempenho, todas as três categorias de granularidade são combinadas projeto algorítmico.

Ferramentas a serem usadas para codificar aplicativos paralelos

As aplicações paralelas que serão desenvolvidas serão escritas em linguagem C.

Serão utilizadas três bibliotecas/APIs específicas para implementar a granularidade paralela apresentada pelos algoritmos

- *biblioteca pthread* – para criar aplicativos multithread para serem executados em ambientes compartilhados arquiteturas de memória (paralelismo de granulação média)
- *MPI (interface de passagem de mensagens)* – para criar aplicações multiprocessadas para serem executadas em arquiteturas de memória distribuída (paralelismo de granulação grossa)
- CUDA C – para criar aplicações onde o paralelismo é expresso em nível variável, destinadas a serem executadas em arquiteturas heterogêneas CPU-GPU (paralelismo de granulação fina).

Leitura sugerida

• *Introdução à HPC com MPI para Ciência de Dados,* Nielsson F., Springer International, 2016
– Capítulo 1: *Uma visão geral da computação de alto desempenho (HPC)* • *Programação de processadores massivamente paralelos: uma abordagem prática,* Kirk DB, Hwu WW, 3ª edição, Morgan Kaufmann, 2017
- Capítulo 1 *Introdução*