VII Simepe
SIMPÓSIO DE ENSINO, PESQUISA E EXTENSÃO
"Expansão e consolidação da Rede Federal: conquistas passadas e perspectivas futuras do IF Sudeste MG"
28 a 30 de janeiro de 2025 | Campus Juiz de Fora

Análise da arquitetura da GPU NVIDIA H100 sob a perspectiva da Arquitetura de Computadores: Inovações e Impactos Acadêmicos

Louzada, V. E. O[1] Garcia, J. S[2] Franco, J. F. S[3] Tavares, R. F[4]

1 Lima, S. M. B, - Campus Rio Pomba – IF Sudeste MG

Introdução

Este estudo aprofundou se na análise da arquitetura da NVIDIA H100 sob a ótica da disciplina de Arquitetura de Computadores, alinhado às diretrizes da Sociedade Brasileira de Computação (SBC) e o livro Arquitetura de Computadores - Uma Abordagem Quantitativa (PATTERSON, HENNESSY, 2013), a partir de uma perspectiva histórica. Explorando a evolução das GPUs e o papel crucial da NVIDIA neste processo. Em seguida, detalhando as características técnicas da H100 com seus diferenciais em relação às gerações anteriores e seu impacto em diversas áreas, como pesquisa científica, desenvolvimento de produtos e serviços baseados em IA,

Objetivo

Explorar a macro e micro arquiteturas:

- Organização dos componentes, exemplo de placa mãe e seus principais componentes, datapath, control path e addres path e pipeline.
- Abordar aspectos básicos da linguagem de montagem (assembly) e exemplos de utilização, bem como uma comparação com o assembly MIPS.

Materiais e Métodos

O artigo foi desenvolvido principalmente por revisão da literatura e o seu espelhamento no conteúdo da disciplina Arquitetura de Computadores conforme as diretrizes da SBC, foram abordados aspectos históricos, arquiteturas macro e micro, e a linguagem de montagem (assembly), comparando-a com o MIPS. Discutiu-se a organização de componentes da placa mãe, caminhos de dados e controle, além de paralelização via pipeline. Também foi comparado o desempenho da GPU com outras CPUs.

Resultados

Apesar de a linguagem assembly não ser acessível publicamente, a H100 oferece um conjunto de instruções semelhante ao MIPS e prioriza o uso de linguagens de alto nível, para aplicações complexas. Sua microarquitetura inclui Tensor Cores otimizados para deep learning, registradores avançados e uma hierarquia de memória eficiente, que juntos maximizam o desempenho.

O pipeline da GPU é altamente paralelo, empregando técnicas para reduzir latências. Componentes como NVLink e NVSwitch otimizam comunicações entre GPUs em configurações de data centers. Em benchmarks, a H100 destaca-se em tarefas paralelas, como treinamento de redes neurais, mas apresenta menor eficiência em operações sequenciais, evidenciando sua especialização em processamento massivo.

Conclusões

A NVIDIA H100 é um marco na evolução das GPUs, desempenhando papel crucial na computação de alto desempenho e IA. Seus avanços em paralelização e precisão tornam uma ferramenta necessária para o processamento de dados em grandes escalas, destacando-se no avanço tecnológico, e como um novo padrão na computação científica e gráfica. Do ponto de vista acadêmico, se propõe o uso da GPU H100 como material didático no ensino de Arquitetura de Computadores, criando um paralelo com a arquitetura MIPS, frequentemente utilizada em cursos introdutórios.

Agradecimentos

Agradeço ao Programa de Educação Tutorial - PET Conexões Ciência da Computação

Apoio: