Google TPU

Fernando Lima
Isabella Caselli
Rodrigo Michelassi

2024

Conteúdo

Resumo

Na era do desenvolvimento de sistemas baseados em Inteligência Artificial e redes neurais profundas (Deep Learning), se faz necessário o uso de máquinas super potentes, capazes de processar dados e realizar operações matemáticas de maneira extremamente rápida. Modelos de Machine Learning podem levar horas, até mesmo dias, para serem treinados, devido principalmente a operações como produto interno entre matrizes e a enorme quantidade de dados que são usados, trazendo um prejuízo não apenas de tempo, mas também energético, ambiental e sobretudo lucrativo. Nesse artigo, iremos tratar brevemente sobre a utilização de Cloud TPUs, unidades de processamento de tensores do Google Cloud, que atuam na otimização do treinamento de modelos de aprendizado de máquina, e que se tornou indispensável na academia e na indústria, para todos estudiosos e profissionais da área.

Capítulo 1

Introdução

A evolução tecnológica tem impulsionado avanços significativos no campo da inteligência artificial, especialmente no desenvolvimento de redes neurais profundas (Deep Learning) e aprendizado de máquina (Machine Learning). Com a crescente complexidade dos modelos e o aumento exponencial na quantidade de dados processados, surgiu a necessidade de hardware especializado, capaz de atender à demanda por poder computacional de maneira eficiente. Nesse contexto, destaca-se a Unidade de Processamento Tensorial (TPU), um circuito integrado de aplicação específica (ASIC) desenvolvido pelo Google para atender às demandas específicas do aprendizado de máquina e inteligência artificial.

Lançadas em 2015, as TPUs foram projetadas para acelerar operações matemáticas essenciais em aprendizado de máquina, como as multiplicações e somas de matrizes, que são fundamentais para o funcionamento de redes neurais. Diferentemente das CPUs e GPUs, cuja arquitetura é generalista e voltada para uma ampla variedade de aplicações, as TPUs possuem uma arquitetura altamente otimizada para tarefas específicas de aprendizado profundo, oferecendo maior eficiência energética e desempenho superior em comparação com soluções tradicionais.

Além de desempenharem um papel crucial no treinamento e na inferência de modelos de aprendizado de máquina, as TPUs também têm permitido avanços significativos em áreas como processamento de linguagem natural, visão computacional e sistemas de recomendação. Sua adoção em larga escala em data centers e projetos de pesquisa reflete a importância desse hardware na computação moderna.

Diante desse cenário, este trabalho tem como objetivo explorar a arquitetura, os princípios de funcionamento e as aplicações das TPUs, destacando sua relevância para os avanços tecnológicos no campo da inteligência artificial e seu impacto na evolução da computação contemporânea.

Capítulo 2

História

A Google Tensor Processing Unit (TPU) é uma arquitetura de circuito integrado especializada, conhecida como um "acelerador de IA". Essa tecnologia teve sua primeira aplicação em 2015, sendo apresentada ao público em 2016 como uma alternativa inovadora às arquiteturas amplamente utilizadas para acelerar o treinamento de modelos de inteligência artificial, como GPUs e arrays sistólicos.

Embora tenha sido oficialmente divulgada apenas em 2016, engenheiros da Google revelaram que a TPU já era utilizada internamente em data centers da empresa por mais de um ano antes de seu lançamento público. Desenvolvida com foco em desempenho otimizado, a arquitetura foi projetada para integrar-se perfeitamente à biblioteca de aprendizado de máquina TensorFlow, também criada pela Google. O TensorFlow é amplamente utilizado para o treinamento de modelos baseados em redes neurais e, atualmente, figura entre as bibliotecas mais populares e robustas do setor.

Uma das principais funcionalidades das TPUs é o processamento eficiente de tensores, uma estrutura de dados representada como um array multidimensional. Estratégias para multiplicação de tensores, tanto de forma linear quanto paralela, são amplamente estudadas em campos como álgebra linear e computação paralela. As TPUs utilizam conceitos avançados dessas áreas para implementar arquiteturas de processamento de produtos matriciais altamente otimizadas, resultando em um desempenho superior em tarefas específicas de aprendizado profundo.

Atualmente, as TPUs são uma tecnologia proprietária da Google, com acesso restrito principalmente aos serviços da empresa. Usuários podem utilizá-las por meio da plataforma de computação em nuvem da Google ou através do Google Colab, onde é possível acessar versões limitadas ou alugar TPUs mais potentes conforme a necessidade.

2.1 Tensores

2.2 Aprendizado de Máquina e Redes Neurais

Em uma era marcada pela presença de modelos como o ChatGPT, IA generativa, e questões éticas envolvendo o uso de imagens e dados pessoais por grandes empresas, surge a necessidade de compreender o funcionamento dos modelos de aprendizado de máquina. Essa compreensão é fundamental não apenas para o fortalecimento científico, mas também para uma aplicação ética e responsável dessas tecnologias. O surgimento das TPUs está diretamente associado ao avanço científico e comercial dessas tecnologias, que dependem cada vez mais de grandes volumes de dados para seu treinamento e operação. Esses modelos, ao lidarem com dados em larga escala, exigem imenso poder computacional para processá-los e realizar cálculos matemáticos intensivos, como produtos tensoriais, fundamentais para redes neurais profundas. De forma resumida, os modelos de aprendizado de máquina consistem na modelagem matemática de problemas com propósitos variados, mas que compartilham um objetivo em comum: permitir que o modelo se ajuste suficientemente bem para fornecer respostas adequadas às entradas recebidas. Em essência, há semelhanças com a computação clássica, onde uma entrada é processada para gerar uma saída correspondente. No entanto, os modelos de aprendizado de máquina possuem uma natureza probabilística, o que significa que não há garantia de que as respostas serão sempre corretas. Ainda assim, são extremamente úteis para resolver problemas difíceis de automatizar. Por exemplo, para um ser humano, identificar a imagem de um gato é uma tarefa trivial. Contudo,

como automatizar tal tarefa? Escrever um algoritmo clássico que reconheça um gato em imagens de diferentes tamanhos, formatos e padrões de cores seria inviável devido à complexidade e variabilidade envolvidas. Nesse cenário, o uso do aprendizado de máquina surge como uma solução poderosa. Os modelos mais modernos de aprendizado de máquina, em especial aqueles baseados em redes neurais, são construídos a partir de duas etapas principais: o forward propagation e o back-propagation. Redes neurais consistem em nós que representam as features dos dados e pesos atribuídos a cada uma dessas features. O forward propagation envolve realizar produtos vetoriais entre os pesos e os valores dos nós, seguidos por uma função não linear aplicada à soma ponderada das entradas. Já o back-propagation atualiza os pesos com base no erro calculado, utilizando novamente produtos vetoriais para ajustar o modelo. Essas operações são computacionalmente intensivas, especialmente em redes neurais que podem conter milhares ou até milhões de parâmetros, dependendo de sua complexidade. Para viabilizar tais cálculos, é indispensável utilizar estratégias que combinem paralelismo e algoritmos eficientes, tarefa para a qual as TPUs foram especificamente projetadas. Conforme será abordado adiante, com sua arquitetura otimizada, as TPUs permitem acelerar o treinamento e a inferência de modelos, tornando-os mais acessíveis e práticos no mundo real.

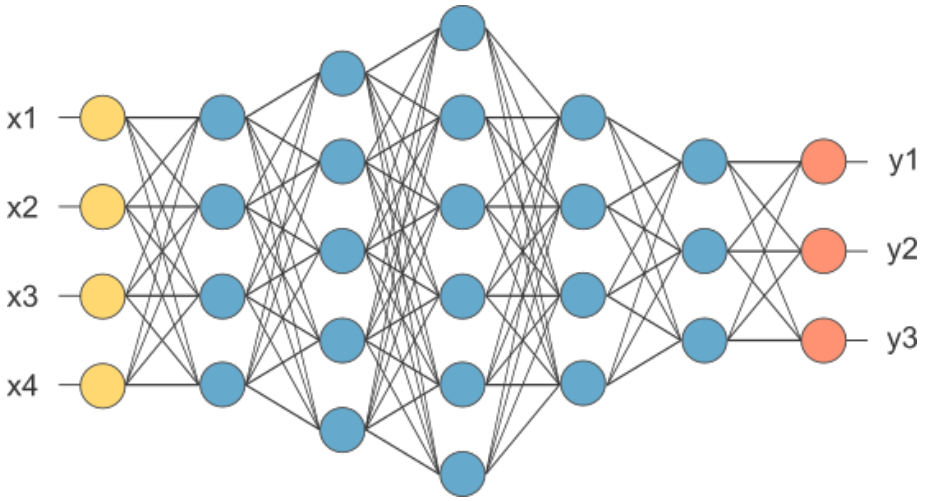

Figura 2.1: Estrutura gráfica de uma Rede Neural

Com base nisso, temos um modelo matemático, que atualiza pesos e é capaz de responder perguntas, porém se ele nem sempre acerta, como eles são de fato utilizados? Vimos que modelos de Aprendizado de Máquina são baseados matemáticos e baseados em pesos. Existem diversas aplicações para esses modelos, e o trabalho de um engenheiro de Machine Learning é conseguir treinar um modelo, com dados o suficiente, de forma a minimizar o erro desse modelo, dessa forma aumentando sua acurácia e garantindo que o modelo tenha uma maior probabilidade de acertar as respostas, com base nos dados que são utilizados como entrada. Atualmente, esses modelos tem diversos usos na indústria, como no mercado financeiro, para predição de séries temporais para variações da bolsa de valores, na medicina para auxílio de profissionais no diagnóstico de doenças, no direito para análise textual de casos, na astronomia, como na recente descoberta da primeira imagem de um buraco negro.

Dessa forma, Machine Learning é uma tendência que cresce tanto no mercado como na academia, e deve continuar sendo explorado por diversos profissionais, de forma que é necessário acelerar o longo processo de treinamento de modelos, além de buscar soluções mais sustentáveis para que o conhecimento possa continuar se expandindo.

Capítulo 3

Arquitetura da TPU

As TPUs surgem como uma alternativa de arquitetura simples para o usuário, oferecendo uma interface de hardware mais amigável e, ao mesmo tempo, capacidade de processar dados com maior rapidez. Essa eficiência foi essencial para atender às demandas de processamento de redes neurais em 2015. Projetada para operar de forma independente da CPU, a arquitetura da TPU recebe instruções diretamente do servidor ao qual está conectada.

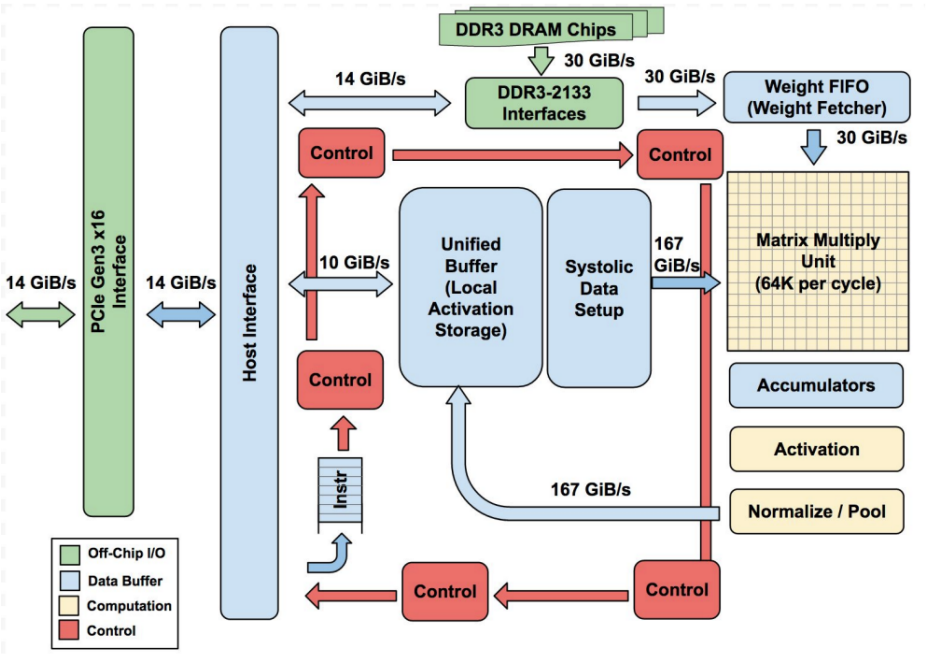

Figura 3.1: Estrutura básica de uma TPU

Para entender melhor a arquitetura de uma TPU, leve em consideração seus pontos mais importantes: a entrada é feita no Weight FIFO, que é levada a principal unidade de processamento, a Matrix Multiply Unit. Essa unidade é responsável por explorar o produto matricial, e iremos explicar mais a fundo em breve. Por fim, a saída é deixada nos acumuladores, e a unidade de ativação performa funções não lineares na saída, para finalmente levar essa saída ao Buffer unificado.

3.1 Unidade de Multiplicação Matricial

Como dissemos, a unidade de multiplicação matricial é o coração da TPU. Essa unidade contém 256×256 MACs (endereço de controle de acesso de mídia) que conseguem processar operações de adição de multiplicação em inteiros (com ou sem sinal) de até 8 bits, gerando produtos de até 16 bits,

armazenados temporariamente nos acumuladores de 32 bits. Esses acumuladores podem carregar até 4MiB de dados. A unidade de processamento de matrizes pode computar até 256 valores por ciclo de clock, e é capaz de realizar operações como produto matricial ou convolução. Vale lembrar que, nos primeiros anos do lançamento da TPU, essa unidade foi planejada justamente para acelerar operações de produto em matrizes densas, sem levar em consideração a esparsidade, que é uma medida útil para acelerar operações de matrizes.

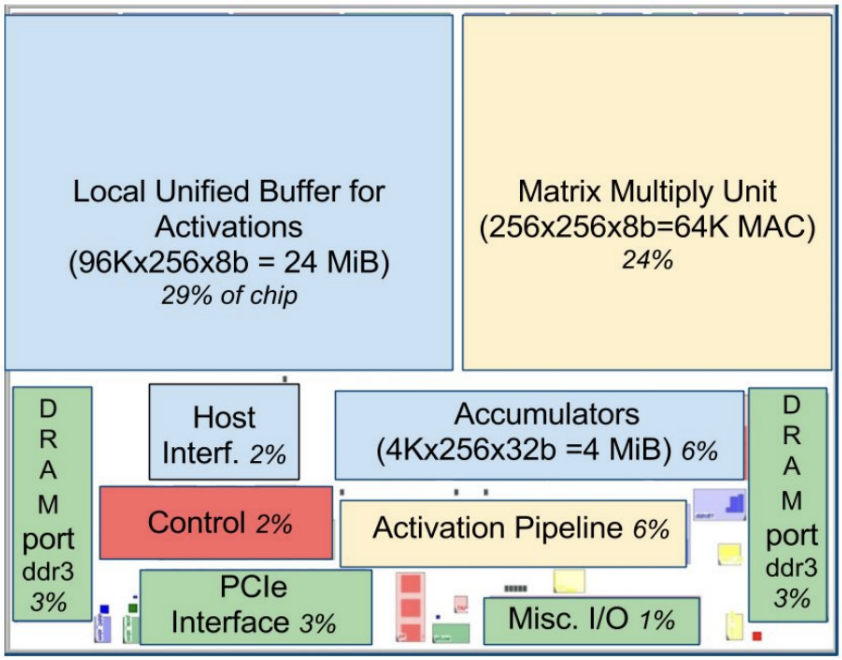

Figura 3.2: Divisão do uso do chip na TPU por unidade

Note que, a unidade de processamento de matrizes irá receber muitos dados de entrada, e ler dados da SRAM consome muito mais energia do que processamento aritmético, portanto essa unidade se aproveita de execução sistólica para economizar energia, reduzindo operações de leitura e escrita do Buffer unificado, dependendo de dados provindos de diversas direções alimentando um array em intervalos regulares.

3.2 Principais instruções

Como vimos, as instruções a serem executadas são enviadas de maneira externa para a TPU, através de PCIe (Peripheral Component Interconnect Express), o que pode ser relativamente lento e caminhar na direção oposta ao que é esperado. Nesse contexto, a fim de acelerar esse processo, a TPU faz uso da arquitetura CISC para definir instruções. A seguir, vemos algumas das principais instruções presentes na TPU:

- Read_Host_Memory: lê dados da CPU para o Buffer Unificado da TPU
- Read_Weights: lê a entrada da unidade matricial
- MatrixMultiply/Convolve: faz com que a unidade matricial performe uma operação de multiplicação ou convolução, e deposite a saída nos acumuladores. Tal operação matricial leva tempo $B \times 256$ da entrada, multiplica por uma constante 256×256 e produz uma saída de tamanho $B \times 256$, e leva B ciclos de clock para ser concluída.
- Activate: performa a função não linear de ativação no neurônio processado que está no acumulador, podendo ser, nativamente, ReLU, Sigmoid, até mesmo pooling para convoluções. Sua saída é levada para o Buffer unificado.
- Write_Host_Memory: escreve os dados do Buffer unificado na CPU.

Como podemos ver, essas principais instruções giram em torno do uso da unidade de processamento de matriz, além de trazer a aplicação de funções conhecidas de Machine Learning, como pooling

e funções de ativação, para serem executadas a nível de hardware, acelerando ainda mais o processo. A filosofia geral da TPU é manter a unidade de matrizes sempre ocupada, e processar diversas instruções paralelamente, para evitar que a unidade de matrizes não tenha entrada assim que acabar uma operação.

3.3 Compatibilidade

Note que a TPU está sendo utilizada sempre de maneira conjunta com a CPU, para recebimento de dados a serem processados. Dessa forma, a pilha de software da TPU tinha como necessidade padrão ser compatível com softwares feitos para rodar na CPU e GPU. Para evitar problemas de compatibilidade, a aplicação de roda na TPU é feita utilizando TensorFlow e compilada em uma API que é capaz de rodar na CPU e GPU.

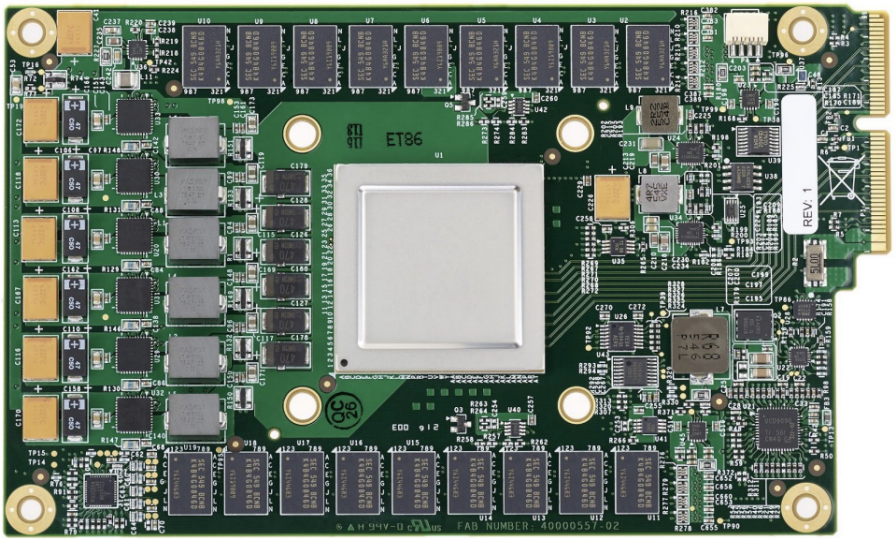

Figura 3.3: Placa de circuito da TPU

Capítulo 4

TPU vs GPU

Capítulo 5

TensorFlow

Capítulo 6

Cloud TPU v5p

Como discutido anteriormente, os modelos de linguagem de grande escala (LLMs) são atualmente os mais explorados tanto no mercado quanto na academia. Esses modelos dependem de grandes volumes de dados, frequentemente oriundos de toda (ou quase toda) a internet, cuja quantidade cresce exponencialmente. Milhões de novos dados são gerados diariamente, o que torna essencial que o hardware acompanhe esse crescimento, oferecendo poder computacional suficiente para treinar modelos dessa magnitude.

Com esse cenário em mente, a evolução das TPUs tem ocorrido de forma acelerada. Em 2023, a Google apresentou a TPU v5p, considerada a mais poderosa já desenvolvida pela empresa. Segundo a multinacional, o novo modelo oferece uma performance até 2,8 vezes mais rápida em comparação às versões anteriores, sendo utilizado para sustentar ecossistemas internos como YouTube, Android e Gmail.

A TPU v5p destaca-se por priorizar desempenho máximo, mesmo que isso sacrifique a facilidade operacional. Um de seus diferenciais está na maneira como lida com operações em ponto flutuante. De acordo com dados oficiais da Google, essa arquitetura incorpora 8.960 chips e apresenta três vezes mais memória HBM, proporcionando um avanço significativo em capacidade de processamento.

	v4	v5e	v5p
Chips por unidade	4.096	256	8.960
Largura de Banda Interconectores	2.400 Gbps	1.600 Gbps	4.800 Gbps
BF16 TFLOPS	275	197	459
Memória HBM	32 GB	16 GB	95 GB
Largura de Banda HBM	1.228 GB/s	820 GB/s	2.795 GB/s

Figura 6.1: Comparativo entre TPUs Google para cargas de trabalho em IA e LLM

Mas no fim, o que significam esses valores? A Google promete que a nova TPU consegue escalar o tempo de treinamento, sendo até $4X$ mais rápida que TPUs mais baratas, como a $V4$, devido ao dobramento no tamanho de operações em ponto flutuante que são entregues.

Capítulo 7

Consumo de Energia

Atualmente, TPUs são encontradas com maior frequência em data centers, e são um produto extremamente utilizado não só pela Google, como também por empresas e usuários que dependem de seus serviços de nuvem.

Para avaliar, primeiramente, a performance/watt da TPU, e comparar com outras arquiteturas como GPU, se utiliza as métricas "total"(que calcula a energia consumida pela TPU/GPU juntamente a CPU, quando calcula performance/watt) e "incremental"(que subtrai a energia consumida pela CPU). A figura abaixo ilustra bem os resultados obtidos pela Google, que mostram uma performance de 17 a 34 vezes melhor que a GPU no modelo total, e de 25 a 29 vezes melhor que a GPU no modelo incremental.

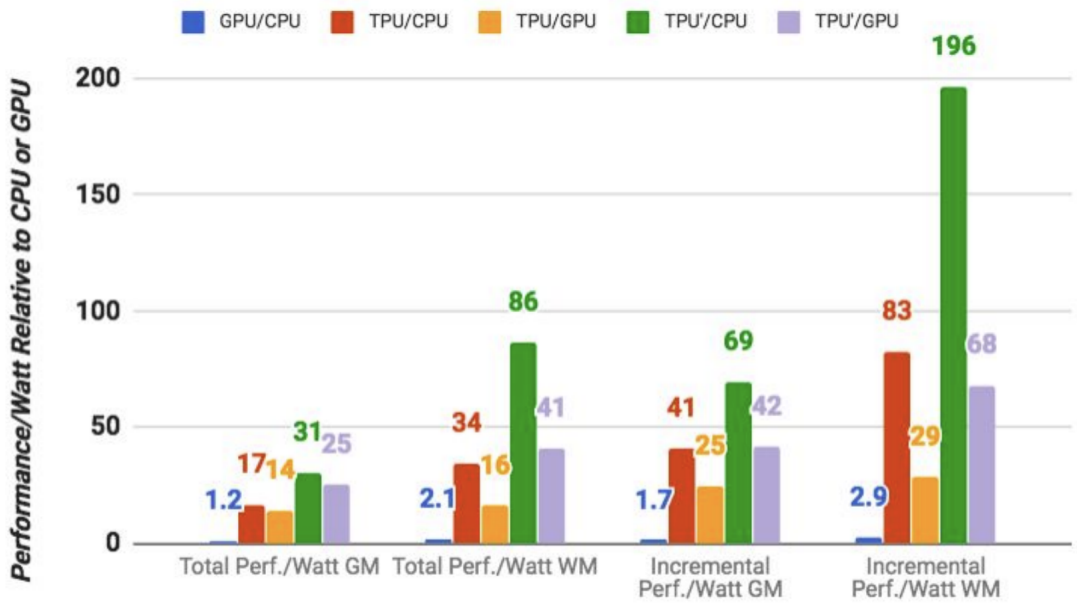

Figura 7.1: Performance/Watt para operações de médias geométricas e ponderadas (GM e WM, respectivamente)

Tendo esses consumos em vista, uma das medidas adequadas para se medir juntamente ao surgimento de uma nova arquitetura é o TDP (Thermal Design Power), que afeta diretamente o custo de energia gasto, tendo em vista que é necessário que a unidade receba energia e resfriamento suficientes. Tendo em vista que servidores não estão em funcionamento durante 100% do tempo, é interessante que a energia consumida por essas máquinas seja proporcional ao seu tempo de uso.

A fim de medir tal consumo e avaliar a validade energética do uso de TPUs em detrimento de GPUs, a figura a seguir compara a energia gasta pelo servidor dividido pelo número de dias, variando o workload para processar CNN0, utilizando o mesmo batch para todos os testes.

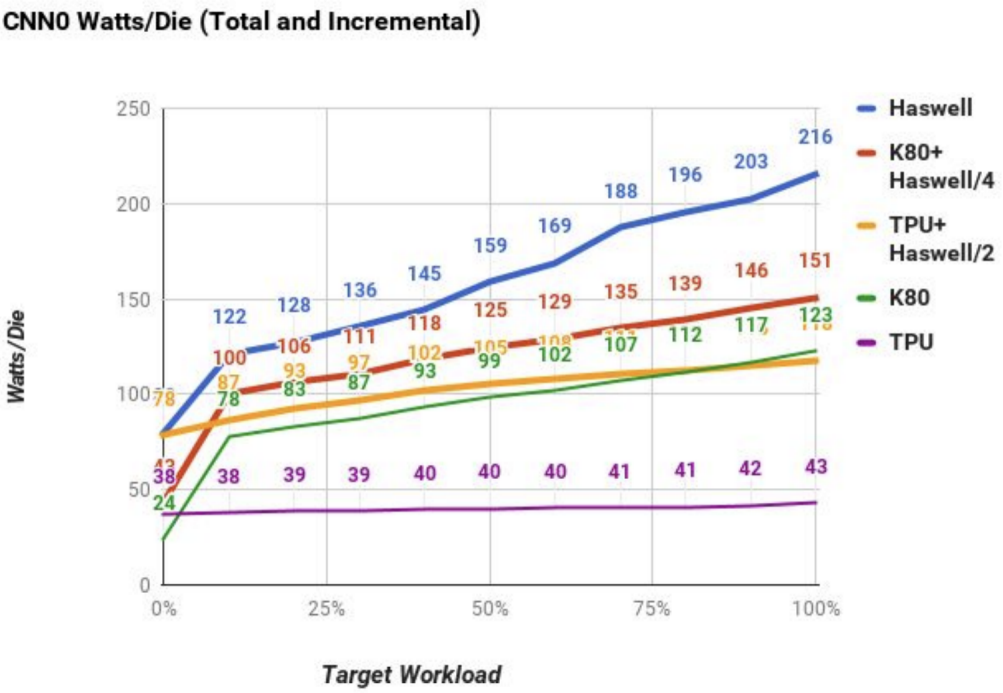

Figura 7.2: Performance/Watt para operações de médias geométricas e ponderadas (GM e WM, respectivamente)

Pelos resultados obtidos, vemos que as TPUs conseguem consumir, no geral, menos energia que as GPUs, porém há um trade-off considerável, pois a quantidade de energia consumida, mesmo com pouca operação na máquina, ainda é muito similar ao consumo com a máquina funcionando em 100%. Ou seja, quando a TPU e GPU estão totalmente carregadas, o servidor de CPU gasta 52% da energia da GPU e 69% da energia da TPU. Nesse sentido, a TPU ainda gasta mais energia proporcionalmente, pois realiza muito mais tarefas que a GPU, porém seus gastos em valor absoluto ainda são inferiores, e justificam o uso desse hardware.

Capítulo 8

Considerações Finais

Bibliografia

TPU e Arquitetura
Comparativos oficiais em modelos de IA (TPU vs GPU)
TPU v5p
Conceitos básicos
Wikipedia, tem tudo
Paper original da TPU