Desempenho e Consumo Energético em Arquiteturas de CPU Homogêneas e Híbridas Utilizando Benchmarks NAS

Yuri Theodoro da Silva, Bruno dos Santos, Vinícius Garcia Pinto

Centro de Ciências Computacionais
Universidade Federal do Rio Grande (FURG)
Rio Grande – RS – Brasil

{yuritheodorods, bruno.santos164439}@gmail.com
vinicius.pinto@furg.br

Resumo. *Este trabalho visa analisar o impacto das tecnologias SMT e Turbo Boost no desempenho e eficiência energética de processadores com arquitetura híbrida (Intel 14ª geração) comparados a uma arquitetura homogênea. Foi usado o conjunto NAS Parallel Benchmarks para mensurar tempo de execução e consumo de energia. Os resultados indicam que a arquitetura híbrida apresenta desempenho superior, impulsionado pelos P-Cores com Turbo Boost, mas sofre com escalabilidade não-linear devido à heterogeneidade dos núcleos. Já a homogênea exibiu escalabilidade consistente até o limite físico dos núcleos. Quanto ao consumo de energia, o Turbo Boost reduz o tempo de execução porém ele eleva a potência instantânea, diminuindo a eficiência em cargas longas.*

1. Introdução

A busca incessante por maior desempenho computacional e eficiência energética tem impulsionado uma evolução constante tanto nas arquiteturas de CPU quanto nas técnicas de programação paralela. Para validar a eficiência dessas aplicações paralelas e das arquiteturas onde elas executam, métricas padronizadas como o NAS Parallel Benchmarks (NPB) [Bailey et al. 1991] destacam-se como referência entre a comunidade científica.

Mais recentemente, o cenário das arquiteturas x86 tornou-se mais complexo com a introdução da arquitetura híbrida da Intel (a partir da 12ª geração). Essa arquitetura combina dois tipos distintos de núcleos: os P-cores (*Performance-cores*) projetados para alto desempenho e baixa latência, e os E-cores (*Efficient-cores*) otimizados para eficiência energética e tarefas de segundo plano. Essa abordagem impõe novos desafios para o gerenciamento de carga de trabalho, exigindo uma compreensão aprofundada de como as tecnologias clássicas (Simultaneous Multithreading – SMT, Turbo Boost, OpenMP) interagem com essa nova assimetria. A introdução de arquiteturas híbridas em processadores de alto desempenho (desktop e servidores) representa uma mudança de paradigma em relação às arquiteturas homogêneas tradicionais. Embora sistemas assimétricos (como *big.LITTLE*) sejam comuns em dispositivos móveis, seu comportamento em cargas de trabalho de Computação de Alto Desempenho utilizando a arquitetura x86 ainda requer caracterizações detalhadas. O objetivo deste trabalho é compreender o impacto prático das tecnologias de paralelismo nesse novo ambiente heterogêneo. Entender a relação entre desempenho e consumo energético em P-Cores e E-Cores é fundamental para desenvolvedores que buscam otimizar suas aplicações. Além disso, a comparação direta com arquiteturas homogêneas tradicionais permite quantificar os reais benefícios e as possíveis limitações da abordagem híbrida para aplicações científicas.

2. Trabalhos Relacionados

A literatura destaca a análise da eficiência energética em arquiteturas multicore por meio de benchmarks NAS. [Shahid et al. 2021] propõem modelos preditivos baseados em contadores de desempenho (PMCs) para otimização energética em sistemas multicore homogêneos. No contexto de hardware híbrido, [Rocha 2025] compara configurações homogêneas e heterogêneas, demonstrando que o escalonamento OpenMP impacta o desempenho e que o uso de E-cores pode maximizar a economia de energia. Complementarmente, [Moori et al. 2025] utilizam o algoritmo genético LOKI para ajuste dinâmico de threads e frequência em processadores Alder Lake, reduzindo o Energy-Delay Product (EDP) em até 85,74% nas aplicações do NPB.

3. Metodologia

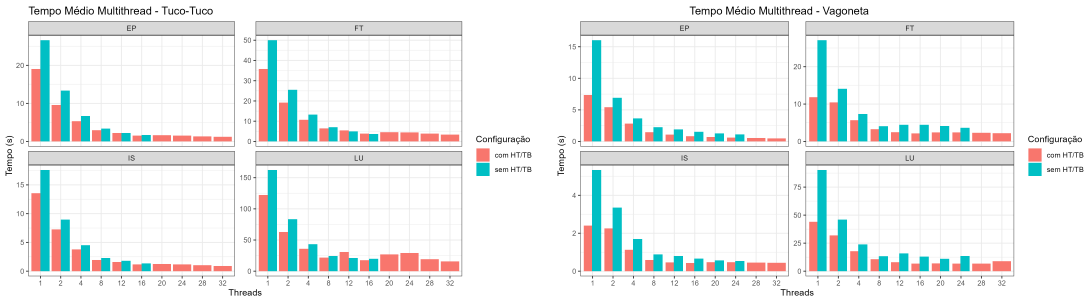
Foram utilizadas as aplicações *Embarrassingly Parallel* (EP), *Lower-Upper Symmetric Gauss-Seidel Solver* (LU), *Integer Sort* (IS) e *Fast Fourier Transform* (FT), todas do NPB. A escolha dessas aplicações permite avaliar distintos padrões de acesso à memória, sincronização entre *threads* e perfis de carga. O EP fornece um *baseline* de cálculo intensivo; o IS estressa contenção no acesso a estruturas de dados compartilhadas; FT explora acessos não contíguos à memória; e LU adiciona relações de dependência de dados entre iterações/*threads*. A coleta de dados energéticos foi realizada através da ferramenta `cpu-energy meter` [Software Systems Lab - LMU Munich 2016], que utiliza a interface RAPL (*Running Average Power Limit*) para registrar o consumo em Joules com precisão. A partir dessas medições, foram avaliadas as seguintes métricas: consumo de energia singlethread, consumo de energia multithread, energia por tempo singlethread, energia por tempo multithread, tempo médio singlethread e tempo médio multithread. Isso permite a comparação não só do ganho de performance proporcionado pelo paralelismo, mas também da eficiência energética em cada cenário. Por limitações de espaço, apresentamos os resultados dos experimentos multithread, reportando tempo de execução e consumo de energia em Joules para ambas as máquinas.

Os experimentos foram executados nas máquinas Tuco-Tuco que possui dois processadores Intel Xeon E5-2640v3, ambos com 8 núcleos homogêneos e hyperthreading e Vagoneta, que possui um processador Intel 14900KF com 8 P-cores e 16 E-cores, com hyperthreading nos P-Cores. Daqui em diante, adotamos a sigla HT para nos referirmos a tecnologia hyperthreading que é uma implementação da Intel para SMT. Todos os núcleos em ambas as máquinas possuem suporte a tecnologia Turbo Boost, referida como TB. As execuções foram feitas utilizando o compilador GCC com suporte ao OpenMP. Para garantir resultados mais confiáveis, foram executadas 30 repetições de cada caso, sendo calculados média e desvio padrão.

4. Resultados

Na Figura 1 temos os gráficos de tempo médio *multithread*. Na máquina Tuco-Tuco, observamos que a execução sem o uso das tecnologias TB e HT começa gastando muito mais tempo, porém conforme o número de *threads* aumenta o tempo de ambas as versões diminui e essa diminuição acontece até oito *threads*, e após continua reduzindo de forma menos agressiva. Em todos os *benchmarks* o tempo com tais tecnologias ativadas é menor. Nota-se que, para execuções com mais de 16 *threads*, a redução do tempo se torna menos

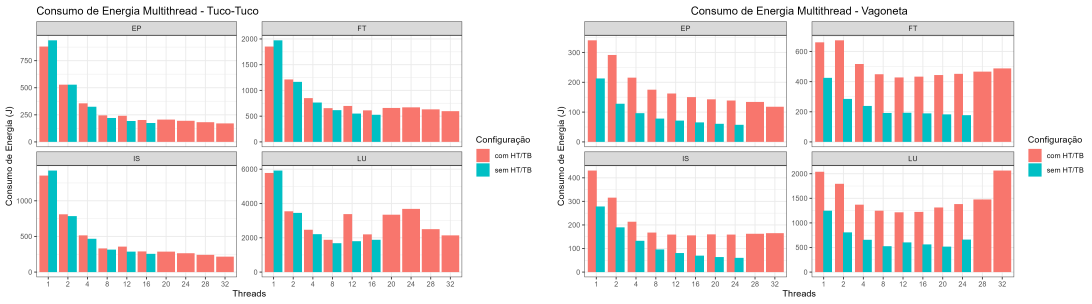
significativa, pois passam a ser utilizados núcleos lógicos, o que pode aumentar o custo de sincronização e a contenção por recursos quando duas *threads* compartilham o mesmo núcleo físico. Já na máquina Vagoneta é possível observar um tempo médio muito maior na execução que não utilizou das tecnologias HT e TB, com a diferença caindo cada vez mais, até ficarem bem próximas a partir da execução com 16 *threads*, o que pode indicar saturação. Isso evidencia que a arquitetura híbrida escala muito bem até o limite dos P-Cores, com o ganho marginal do paralelismo sendo reduzido a partir da entrada dos E-cores. Na Tuco-Tuco a escalabilidade é linear e previsível até o limite de seus núcleos físicos. Já a Vagoneta, atinge tempos absolutos bem menores com poucas *threads* devido à força dos P-Cores, porém enfrenta dificuldades de escalabilidade mais cedo.



(a) Tuco-Tuco **(b) Vagoneta**

Figura 1. Tempo Médio das Execuções com Múltiplas *Threads* usando ou não as tecnologias Hyperthreading (HT) e Turbo Boost (TB).

Na Figura 2 temos o consumo total de energia em Joules. Na Tuco-tuco, a execução sem as tecnologias HT e TB consome mais energia com uma única *thread*, mas isto se inverte ao usar mais *threads*. É possível notar que, conforme o número de *threads* aumenta, o consumo de energia diminui até chegar nos núcleos lógicos, onde continua diminuindo, porém em uma escala muito menor. Isso mostra que cargas de trabalho paralelas podem ser energeticamente eficientes, desde que a escalabilidade seja boa e o *overhead* seja baixo. Na máquina híbrida Vagoneta, diferente da Tuco-Tuco, a execução com HT e TB consome mais energia em todos os casos, com o consumo diminuindo de maneira agressiva com o aumento do número de threads em ambos os casos até aproximadamente 12/16 *threads*, o que indica que o paralelismo é energeticamente eficiente. O aumento a partir das 16 *threads* se deve ao uso do HT e ativação agressiva do TB, pois isso eleva a potência e a energia volta a subir. Isso mostra que a execução sem as tecnologias tende a ser mais eficiente energeticamente e que paralelizar moderadamente é energeticamente vantajoso, pois reduz o consumo total de energia, mas nem sempre o uso das tecnologias é a melhor opção, pois os ganhos de desempenho podem não compensar o custo energético adicional. A comparação do consumo total permite observar uma dinâmica de eficiência distinta. A Vagoneta é econômica em baixas contagens de *threads*, superando a Tuco-Tuco com larga margem. Porém, em carga máxima de 32 *threads*, essa vantagem diminui bastante. Isso acontece porque o custo energético para ativar todas as tecnologias da Vagoneta eleva o consumo total, fazendo com que a energia gasta volte a subir em *benchmarks* com FT e LU. Enquanto isso, a Tuco-Tuco mantém um padrão de consumo mais linear, chegando até mesmo a empatar com o consumo em cenários pesados como o LU, onde a Vagoneta perde seu posto de mais eficiente.



(a) Tuco-Tuco **(b) Vagoneta**

Figura 2. Energia Total Multithread. Comparação entre execuções com ou sem as tecnologias Hyperthreading (HT) e Turbo Boost (TB).

5. Conclusões

Este trabalho analisou o impacto das tecnologias *Hyperthreading* e *Turbo Boost* no desempenho e consumo energético, comparando uma arquitetura homogênea com uma híbrida. Executando *benchmarks* do NPB foram quantificadas diferenças nas arquiteturas quanto ao consumo energético e tempo de execução sob cargas diferentes de trabalho.

O uso destas tecnologias mostrou-se eficaz para a redução do tempo médio de execução em ambas as arquiteturas, com reduções que passaram de 50% com EP e FT. Porém, esse ganho de desempenho cobra um alto preço na eficiência energética. Na Vagoneta, a ativação dessas tecnologias elevou muito a potência média e, em muitos casos, o consumo total de energia, evidenciando um "trade-off" com o desempenho máximo sendo atingido às custas de uma menor eficiência energética. Em relação à escalabilidade a máquina Tuco-Tuco apresentou um comportamento previsível e linear até o limite de seus núcleos físicos, enquanto a Vagoneta mostrou-se excelente em escalabilidade nos P-Cores, porém sofreu com ganhos marginais decrescentes e perda de eficiência ao utilizar os E-Cores em tarefas que dependiam muito de memória.

Como trabalhos futuros pretende-se estender a análise para outros *benchmarks* e aplicações, analisar isoladamente a influência do Turbo Boost e do Hyperthreading e explorar o uso dos E-Cores de maneira isolada. Além disso, também está prevista a análise de arquiteturas assimétricas de outros fabricantes.

Referências

Bailey, D., Barszcz, E., Barton, J., et al. (1991). The NAS parallel benchmarks. *Int. J. High Perform. Comput. Appl.*, 5(3):63–73.

Moori, M. K., Rocha, H. M. G. D. A., et al. (2025). Energy-efficient execution of parallel regions on heterogeneous multi-cores. In *Proceedings of the 38th SBC/SBMicro*. SBC.

Rocha, R. B. (2025). Analysis of openmp scheduling policies in hybrid architectures.

Shahid, A., Fahad, M., Manumachu, R. R., et al. (2021). Energy predictive models of computing: theory, practical implications and experimental analysis on multicore processors. *IEEE Access*, 9:84675–84693.

Software Systems Lab - LMU Munich (2016). CPU energy meter. `https://github.com/sosy-lab/cpu-energy-meter`. Acessado em: 6 jun.2025.