

Maximizando Ciência por Joule com Precisão Numérica e DVFS em GPUs

Átila Uebel¹, Matheus Costa¹, Arthur F. Lorenzon¹

¹Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{auebel, matheus.costa, aflorenzon}@inf.ufrgs.br

Resumo. A eficiência energética em computação científica deve considerar não apenas a potência consumida, mas também a quantidade de trabalho útil concluído por unidade de energia. Neste trabalho, investigamos como a precisão numérica afeta essa relação na dinâmica molecular acelerada por GPU. Avaliamos três cargas do LAMMPS, com diferentes participações de cálculos de longo alcance, em três GPUs NVIDIA Ada Lovelace, considerando FP64, FP32/64 e FP32, além de diferentes frequências de operação. Os resultados mostram que o benefício da precisão reduzida depende fortemente da carga e da plataforma: em FP32, a redução de EDP varia de 8,1% a 88,6%. Quando o maior throughput também reduz o tempo de execução, o ganho energético é substancial, alcançando até $3,76\times$ mais atom-steps por joule. Em cargas pouco sensíveis à precisão, esse benefício permanece próximo de $1,1\times$. Mostramos ainda que FP32/64 preserva grande parte dos ganhos e que a frequência que minimiza o EDP varia conforme o modo numérico, indicando que maximizar a eficiência por joule requer considerar conjuntamente a precisão, as características da carga e a frequência de operação.

1. Introdução

O consumo de energia tornou-se uma restrição importante na computação de alto desempenho (HPC – *High-Performance Computing*), com impacto direto nos custos operacionais e na capacidade de execução de grandes centros de processamento [Afzal et al. 2026a]. Nesse contexto, aplicações científicas aceleradas por GPU têm sido avaliadas não apenas pelo tempo de solução, mas também pela quantidade de computação que conseguem realizar dentro de um determinado orçamento energético [Godoy et al. 2026]. Entre essas aplicações está a dinâmica molecular (DM), utilizada para simular a evolução temporal de sistemas de partículas e implementada em motores de produção como LAMMPS [Thompson et al. 2022] e GROMACS [Abraham et al. 2015]. Esses códigos exploram amplamente aceleradores de GPU em sistemas HPC atuais [Johansson et al. 2025], tornando relevante identificar configurações que permitam realizar mais trabalho científico com a energia disponível.

Uma possibilidade é reduzir a precisão numérica utilizada na simulação. Modos FP32 e de precisão mista podem explorar o maior *throughput* das unidades de precisão simples e reduzir o volume de dados movimentados em comparação com o FP64. Nas GPUs Ada avaliadas neste trabalho, por exemplo, o pico de processamento em FP32 pode ser até 64 vezes maior do que o de FP64 [NVIDIA Corporation 2023]. Estratégias desse

tipo já são utilizadas por motores de DM como AMBER [Le Grand et al. 2013] e GRO-MACS [Abraham et al. 2015]. Entretanto, a existência de maior capacidade computacional em FP32 não implica que toda simulação apresente o mesmo ganho de desempenho ou de energia. Cargas de DM combinam, em diferentes proporções, cálculos de curto alcance, operações sobre malhas, transformadas e movimentação de dados; à medida que etapas pouco sensíveis ao throughput em FP32 passam a dominar a execução, o benefício da menor precisão pode diminuir substancialmente.

Essa dependência ainda é pouco caracterizada sob a perspectiva energética. Estudos sobre precisão reduzida em DM concentram-se principalmente no desempenho e na fidelidade numérica, enquanto trabalhos sobre eficiência energética investigam mecanismos, como o controle de frequência, geralmente em uma configuração fixa de precisão [Afzal et al. 2026b, Machado et al. 2026]. Portanto, permanece a questão de quando a redução de precisão efetivamente permite concluir mais trabalho científico por unidade de energia, de qual mecanismo esse ganho decorre e como a mudança do modo numérico interfere no ponto energeticamente eficiente de operação da GPU.

Neste trabalho, investigamos essa relação utilizando três cargas do LAMMPS com participações distintas nos cálculos de longo alcance: LJ, SPC/E e NaCl, para as quais a fração do tempo em *k-space* varia de 0% a 93%. Avaliamos três modos numéricos, FP64, FP32 e FP32/64, em três GPUs NVIDIA Ada Lovelace de diferentes classes: L40S, RTX 4070 e RTX 4090. Além do comportamento na frequência padrão, realizamos *sweeps* de frequência para determinar como o modo numérico modifica a região de menor EDP. As comparações entre precisões são realizadas em condições de execução equivalentes, e os modos de precisão reduzida empregados foram previamente validados pelos autores quanto às propriedades físicas e numéricas relevantes das simulações.

Os resultados mostram que o retorno energético da precisão reduzida depende fortemente da carga e da plataforma. Na L40S, por exemplo, FP32 reduz o EDP em 68,8% no LJ, que não utiliza *k-space*, mas em apenas 8,1% no NaCl, no qual essa etapa representa 93% da execução. A decomposição do EDP mostra a origem dessa diferença, em que, no LJ, a menor precisão reduz simultaneamente o tempo de execução e a potência, enquanto, no NaCl, a aceleração praticamente desaparece e resta principalmente uma redução de potência. O SPC/E demonstra ainda que cargas com composição semelhante podem responder de forma distinta entre GPUs, evidenciando uma interação entre o perfil computacional e a arquitetura. A precisão também modifica a frequência que minimiza o EDP, mostrando que precisão e frequência não devem ser tratadas como mecanismos de otimização independentes.

Essa diferença tem uma consequência direta sobre a produtividade energética da simulação. Quando o maior throughput da menor precisão é efetivamente convertido em menor tempo de solução, a mesma quantidade de trabalho científico pode ser realizada com substancialmente menos energia. No LJ, FP32 alcança até $3,76\times$ mais *atom-steps* por joule do que FP64. Em cargas dominadas por etapas pouco sensíveis à precisão, como o NaCl, esse ganho permanece próximo de $1,1\times$. Nesse sentido, mostramos também que a precisão reduzida permite transformar a capacidade computacional disponível em mais ciência realizada por joule.

2. Fundamentação e Trabalhos Relacionados

2.1. Dinâmica molecular

A dinâmica molecular (DM) simula a evolução temporal de sistemas de partículas por meio da integração das equações de movimento, normalmente utilizando métodos derivados de Verlet [Verlet 1967]. Uma parcela importante do custo computacional está associada ao cálculo das forças, que podem envolver interações de curto e de longo alcance. Interações de curto alcance, como o potencial de Lennard-Jones (LJ), são avaliadas dentro de um raio de corte, enquanto interações eletrostáticas de longo alcance são normalmente tratadas por métodos baseados em malhas, como PPPM e PME [Hockney and Eastwood 1988, Essmann et al. 1995]. No LAMMPS, essa última componente é reportada como *kspace* [Thompson et al. 2022].

Essas etapas apresentam características computacionais distintas na GPU. Cálculos de curto alcance tendem a explorar mais diretamente a capacidade de processamento do acelerador, enquanto operações associadas a *kspace* envolvem malhas, transformadas e maior movimentação de dados [Trott et al. 2010]. Como consequência, a fração de tempo dedicada a cada componente pode alterar a sensibilidade da aplicação ao throughput computacional da GPU. Neste trabalho, utilizamos a fração do tempo de execução em *kspace*, obtida diretamente dos logs do LAMMPS, para caracterizar essa mudança de comportamento entre as cargas avaliadas.

2.2. Precisão numérica em dinâmica molecular

Aplicações de DM aceleradas por GPU podem utilizar diferentes representações numéricas. FP64 oferece maior precisão, enquanto FP32 reduz o volume de dados e permite explorar uma capacidade de processamento substancialmente maior em GPUs. A precisão mista procura combinar essas características, utilizando maior precisão apenas nas estruturas mais sensíveis à acumulação de erro. Em DM, essa estratégia já foi utilizada para acelerar os cálculos de forças, mantendo propriedades numéricas adequadas para a simulação [Le Grand et al. 2013].

A redução de precisão, entretanto, não implica necessariamente um ganho uniforme de desempenho. O benefício depende das operações predominantes na execução. Componentes sensíveis ao throughput computacional podem ser fortemente acelerados, enquanto etapas limitadas por outros recursos podem apresentar pouca alteração. Além disso, a escolha da precisão deve preservar propriedades físicas relevantes da simulação, uma vez que erros de arredondamento podem se acumular durante integrações longas [Colberg and Höfling 2011].

Grande parte da literatura sobre precisão reduzida em aplicações científicas concentra-se justamente no compromisso entre desempenho e fidelidade numérica [Kashi et al. 2026]. Sua relação com eficiência energética tem recebido menor atenção. Godoy et al. [Godoy et al. 2026], por exemplo, mostram que diferentes representações numéricas podem alterar o consumo de energia de aplicações científicas em GPUs distintas, mas não investigam a dinâmica molecular nem como esse efeito depende da composição computacional da carga. Nosso estudo parte dessa lacuna e analisa conjuntamente o tempo de execução, a potência, a energia e o EDP, relacionando esses efeitos ao perfil das cargas de DM e, posteriormente, ao trabalho científico por joule.

2.3. Frequência de operação e eficiência energética

A frequência de operação da GPU constitui outro mecanismo para controlar o compromisso entre desempenho e consumo de energia. A redução do *clock* pode diminuir a potência dinâmica, mas também aumentar o tempo necessário para concluir a aplicação. Como resultado, a frequência que minimiza o consumo ou o EDP depende da sensibilidade da carga ao desempenho do processador [Mei et al. 2017]. Aplicações fortemente dependentes de computação tendem a sofrer maior penalidade quando a frequência é reduzida, enquanto outras podem operar em frequências menores com impacto limitado no tempo de execução.

Esse comportamento tem sido explorado em aplicações de DM. O GROMACS *Unplugged* [Afzal et al. 2026b], por exemplo, mostra que a frequência da GPU pode alterar significativamente o compromisso entre desempenho e energia em diferentes plataformas. Trabalhos relacionados também demonstram que o ponto de maior eficiência depende das características da aplicação e da arquitetura utilizada [Machado et al. 2026]. Essas análises, entretanto, consideram essencialmente uma configuração numérica fixa.

Neste trabalho, investigamos uma dimensão adicional: se a própria mudança na precisão altera a resposta da aplicação à frequência. Em vez de tratar a precisão e o *clock* como decisões independentes, avaliamos se o ponto de menor EDP se desloca entre FP64, FP32/64 e FP32. Essa análise complementa a caracterização do efeito precisão $\times$ carga e permite determinar se o maior trabalho científico por unidade de energia também depende da configuração de frequência utilizada.

3. Metodologia

3.1. Plataformas Avaliadas

Os experimentos foram realizados no cluster do Parque Computacional de Alto Desempenho (PCAD) da UFRGS utilizando três GPUs NVIDIA Ada Lovelace: RTX 4070 (12 GB, TDP de 200 W), RTX 4090 (24 GB, TDP de 450 W) e L40S (48 GB, TDP de 350 W). A escolha de dispositivos da mesma microarquitetura permite avaliar GPUs das classes *consumer* e *datacenter* sem introduzir diferenças decorrentes de gerações distintas. O ambiente experimental utiliza CUDA 12.4 e driver NVIDIA 590.44.01. Cada execução utiliza exclusivamente uma GPU, alocada por meio do SLURM.

3.2. Cargas de Trabalho

A avaliação utiliza três cargas do LAMMPS [Thompson et al. 2022] (versão estável de agosto de 2024): **LJ**, um fluido de Lennard-Jones com 108.000 átomos; **SPC/E**, um sistema de água rígida com 108.000 átomos e interações eletrostáticas calculadas por PPPM; e **NaCl**, um sistema de sal fundido com 110.592 átomos e PPPM. Todas as cargas executam 5.000 passos de integração. O número de passos foi definido com base em uma avaliação preliminar entre 10^3 e 10^6 passos, na qual verificamos a estabilização da energia consumida por átomo-passo a partir de 5.000 passos. Essa duração também evita execuções excessivamente curtas para a coleta de potência, em conformidade com as recomendações para medições energéticas em GPUs [Yang et al. 2024].

As três cargas foram selecionadas para representar diferentes participações em cálculos de longo alcance. A partir do *MPI task timing breakdown* reportado pelo LAMMPS,

calculamos a fração do tempo de execução associada a *kspace*: 0% para LJ, aproximadamente 69% para SPC/E e 93% para NaCl. Utilizamos essa medida como caracterização operacional das cargas ao longo do artigo. Dessa forma, a análise não assume a priori que uma carga seja limitada pela computação ou pela movimentação de dados, mas relaciona os efeitos observados de precisão à participação crescente da etapa de longo alcance.

3.3. Configurações Experimentais

Avaliamos três modos de precisão numérica: FP64, utilizado como referência; FP32, no qual as principais estruturas de dados utilizam `float`; e FP32/64, que mantém posições em `double` e utiliza `float` para forças e velocidades. As versões foram geradas por meio dos mecanismos de compilação disponibilizados pelo próprio LAMMPS, incluindo `-DGPU_PREC`, sem modificar os algoritmos da aplicação. Os modos de precisão reduzida utilizados no estudo foram previamente validados pelos autores quanto às propriedades físicas e numéricas relevantes das simulações; portanto, a análise energética considera apenas configurações que preservam os resultados científicos, de acordo com os critérios de validação adotados.

A avaliação é dividida em dois conjuntos experimentais. No primeiro, comparamos FP64, FP32/64 e FP32 na frequência de referência da GPU para quantificar o efeito da precisão sobre o tempo de execução, a potência, a energia e o EDP. No segundo conjunto, realizamos um *sweep* da frequência dos SMs para cada combinação de carga e de precisão. Na L40S, avaliamos oito frequências entre 300 e 2.490 MHz: 300, 600, 900, 1.200, 1.500, 1.800, 2.100 e 2.490 MHz. Nas RTX 4070 e RTX 4090, avaliamos sete níveis: 300, 675, 1.050, 1.425, 1.800, 2.175 MHz e, respectivamente, 2.475 e 2.520 MHz como frequências máximas. O *clock* é configurado antes de cada execução por meio de um *wrapper* institucional em modo de persistência e sua aplicação é verificada pela frequência efetivamente reportada pelo dispositivo. Cada configuração é executada dez vezes, e a mediana das repetições é utilizada para representar o comportamento.

3.4. Coleta de Métricas e Análise

Durante cada execução, coletamos potência e frequência da GPU a 10 Hz por meio da NVML, utilizando `nvidia-smi`. A energia consumida (E) é obtida pela integração de `power.draw.instant` durante a janela efetiva de execução da aplicação, excluindo os períodos anteriores e posteriores à carga computacional. A partir dessas medições, analisamos o tempo de execução (T), a potência média (P), a energia consumida (E) e o *energy-delay product* (EDP).

Para reduzir a influência de diferenças nas alocações do cluster, as comparações entre modos de precisão são realizadas na mesma alocação, utilizando FP64 como referência. Os efeitos de FP32 e FP32/64 são, portanto, expressos como valores relativos de tempo, potência, energia e EDP em relação à execução em FP64 realizada sob as mesmas condições.

Além do EDP, quantificamos a quantidade de trabalho científico realizada por unidade de energia. Como cada execução processa um número conhecido de átomos (N_{atoms}) durante um número fixo de passos (N_{steps}), definimos a eficiência científica como

$$\eta_{\text{sci}} = \frac{N_{\text{atoms}} \times N_{\text{steps}}}{E}, \quad (1)$$

expressa em *atom-steps* por joule. Para comparar os modos numéricos dentro de uma mesma carga e arquitetura, normalizamos essa métrica em relação ao FP64:

$$\eta_{\text{sci,norm}} = \frac{\eta_{\text{sci}}}{\eta_{\text{sci,FP64}}} = \frac{E_{\text{FP64}}}{E}. \quad (2)$$

Valores superiores a 1 indicam, portanto, que uma quantidade maior de trabalho científico é realizada por unidade de energia. Para quantificar formalmente a dependência do EDP em relação à precisão e à carga, aplicamos uma ANOVA fatorial do Tipo III com contrastes *Sum* sobre $\log(\text{EDP})$ [Hoeﬂer and Belli 2015, Fox and Weisberg 2019]. Além da signiﬁcância estatística, reportamos o tamanho de efeito por meio de η_p^2 e seu intervalo de conﬁança de 90%. Todos os dados e análises foram processados em Python utilizando *pandas* e *statsmodels*.

4. Resultados

Nesta seção, analisamos quando a redução de precisão se converte em maior eficiência energética e quais mecanismos determinam esse comportamento. Primeiro, quantificamos seu impacto sobre o EDP e a dependência em relação à carga. Em seguida, separamos os efeitos sobre o tempo de execução e potência. Depois, avaliamos como a precisão afeta a resposta ao ajuste de frequência. Por fim, traduzimos os ganhos energéticos em trabalho científico realizado por joule.

4.1. Impacto da precisão sobre o EDP

O benefício energético da precisão reduzida varia substancialmente entre as cargas avaliadas. A Figura 1 mostra que FP32 apresenta as maiores reduções de EDP no LJ: 68,8% na L40S, 88,6% na RTX 4070 e 73,6% na RTX 4090. No NaCl, em contraste, as reduções são de apenas 8,1%, 15,4% e 10,6%, respectivamente. Assim, uma mesma alteração de precisão pode produzir desde ganhos marginais até reduções superiores a 80% no EDP. Esse contraste acompanha a participação crescente dos cálculos de longo alcance. O LJ não utiliza *kspace*, enquanto essa etapa representa aproximadamente 69% do tempo no SPC/E e 93% no NaCl. Na L40S, por exemplo, o EDP normalizado de FP32 aumenta de 0,31 no LJ para 0,69 no SPC/E e 0,92 no NaCl. À medida que o *kpace* passa a dominar a execução, diminui o retorno do maior *throughput* disponível em menor precisão.

A fração de *kpace*, contudo, não explica, isoladamente, o comportamento. O SPC/E reduz o EDP em 31,4% na L40S e 23,1% na RTX 4090, mas chega a 84,2% na RTX 4070. Isso mostra que o efeito da precisão resulta da interação entre as características da carga e a plataforma em que ela é executada. FP32/64 acompanha essa mesma tendência e preserva aproximadamente 76% a 100% da redução de EDP alcançada por FP32 nas combinações avaliadas. A proximidade entre os dois modos é particularmente evidente nas cargas em que a redução de precisão afeta pouco o desempenho. A interação precisão×carga também é estatisticamente signiﬁcativa, com a ANOVA fatorial Tipo III apresentando $F(4, 126) = 5,86$, $p = 0,0002$ e $\eta_p^2 = 0,157$, e um intervalo de conﬁança de 90% entre 0,05 e 0,23, confirmando que não há um benefício energético único associado à menor precisão.

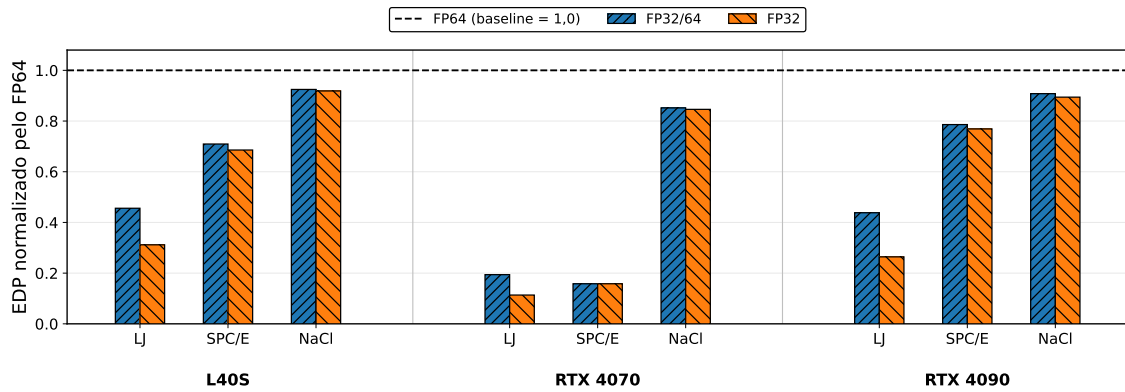


Figura 1. EDP de FP32 e FP32/64 normalizado ao FP64 para as três cargas e arquiteturas avaliadas. Valores inferiores a 1 indicam redução do EDP em relação ao FP64.

4.2. Decomposição do ganho de EDP em tempo e potência

A Figura 2 mostra como o efeito anterior se distribui entre o tempo de execução e a potência média, ambos normalizados ao FP64. Deslocamentos para a esquerda indicam menor tempo de execução, e deslocamentos para baixo indicam menor potência. No LJ, FP32 reduz simultaneamente as duas componentes em todas as três GPUs. O caso mais pronunciado ocorre na RTX 4070, com redução de 56,5% no tempo e 42,2% na potência. Nesse regime, o maior *throughput* de FP32 resulta em menor tempo de solução e, simultaneamente, em menor demanda energética.

O NaCl apresenta um mecanismo distinto, em que o tempo varia em menos de 2,1% entre FP64 e FP32 nas três plataformas, enquanto a potência diminui entre 7,4% e 15,5%. Assim, praticamente não há aceleração na simulação, e o ganho energético decorre predominantemente da redução da potência. Isso explica o EDP próximo ao FP64 observado nessa carga. O SPC/E expõe novamente a dependência da plataforma, uma vez que na L40S e na RTX 4090, FP32 pouco altera o tempo, mas reduz a potência em 27,5% e aproximadamente 20%, respectivamente. Por outro lado, na RTX 4070, tempo

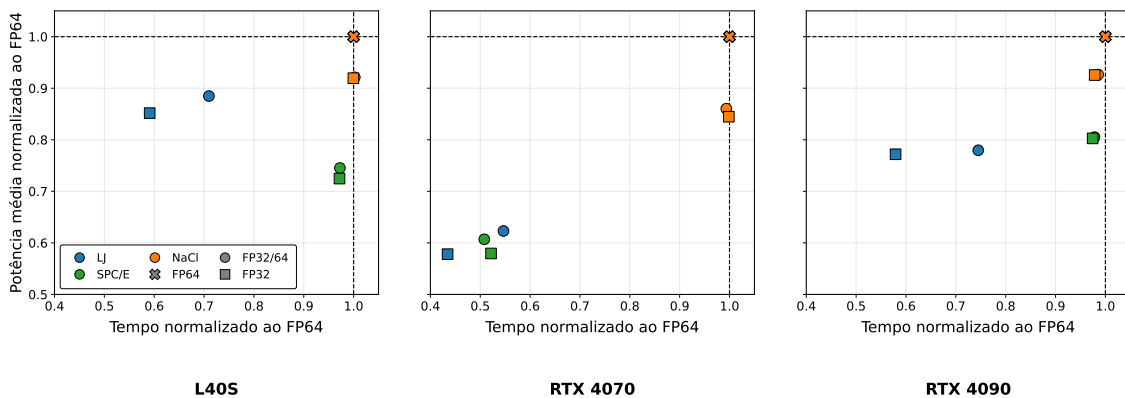


Figura 2. Tempo de execução e potência média normalizados ao FP64. Cores identificam as cargas e marcadores os modos de precisão. Deslocamentos para a esquerda representam menor tempo de execução e deslocamentos para baixo, menor potência média.

Tabela 1. Frequência que minimiza o EDP para cada combinação de carga, GPU e precisão.

GPU	Carga	FP64	FP32/64	FP32
Frequência de menor EDP (MHz)				
L40S	LJ	2100	1800	900
	SPC/E	2100	2100	1800
	NaCl	1500	1200	1200
RTX 4070	LJ	2475	2175	1800
	SPC/E	2475	2175	2175
	NaCl	1800	1800	1800
RTX 4090	LJ	2175	1425	2175
	SPC/E	2175	2175	2175
	NaCl	1425	1425	1050

e potência diminuem em aproximadamente 48% e 42%. Portanto, os maiores ganhos de EDP ocorrem quando a menor precisão atua em ambas as dimensões, reduzindo a potência e, principalmente, permitindo concluir a simulação em menos tempo.

4.3. Interação entre precisão e frequência de operação

A mudança no balanço entre tempo de execução e potência também altera a resposta da aplicação ao DVFS. A Figura 3 apresenta o EDP obtido ao variar a frequência dos SMs, normalizado pelo valor observado na maior frequência avaliada para cada combinação de GPU, carga e precisão. Valores inferiores a 1 indicam, portanto, que reduzir a frequência melhora o EDP em relação à operação no maior *clock* disponível. Os resultados mostram que frequências muito baixas tendem a aumentar rapidamente o EDP, pois a redução de potência deixa de compensar o aumento no tempo de execução. Por outro lado, o maior *clock* também não corresponde necessariamente à configuração energeticamente mais eficiente. Na RTX 4070, o espaço adicional de otimização é relativamente pequeno, com reduções de até aproximadamente 6%. Na L40S e na RTX 4090, algumas combinações apresentam reduções próximas ou superiores a 20%, indicando maior margem para exploração de DVFS.

Mais importante, a frequência que minimiza o EDP depende da precisão adotada. A Tabela 1 apresenta esse ponto para todas as combinações avaliadas. Ao passar de FP64 para FP32, a frequência de menor EDP varia em seis das nove combinações entre carga e GPU. Sempre que ocorre, o deslocamento é para uma frequência inferior à observada em FP64. Entre essas configurações, a redução mediana da frequência de menor EDP é de aproximadamente 23%.

No LJ da L40S, o ponto de menor EDP passa de 2100 MHz em FP64 para 1800 MHz em FP32/64 e apenas 900 MHz em FP32, uma redução de 57% em relação ao ponto observado com precisão dupla. Na RTX 4070, o LJ apresenta comportamento semelhante, com redução de 2475 MHz para 1800 MHz entre FP64 e FP32. Já no NaCl da RTX 4090, a frequência de menor EDP passa de 1425 para 1050 MHz. Esses casos mostram que uma configuração de DVFS definida em FP64 pode deixar de ser adequada após a alteração do modo numérico.

No NaCl da RTX 4070, a frequência de menor EDP permanece em 1800 MHz para

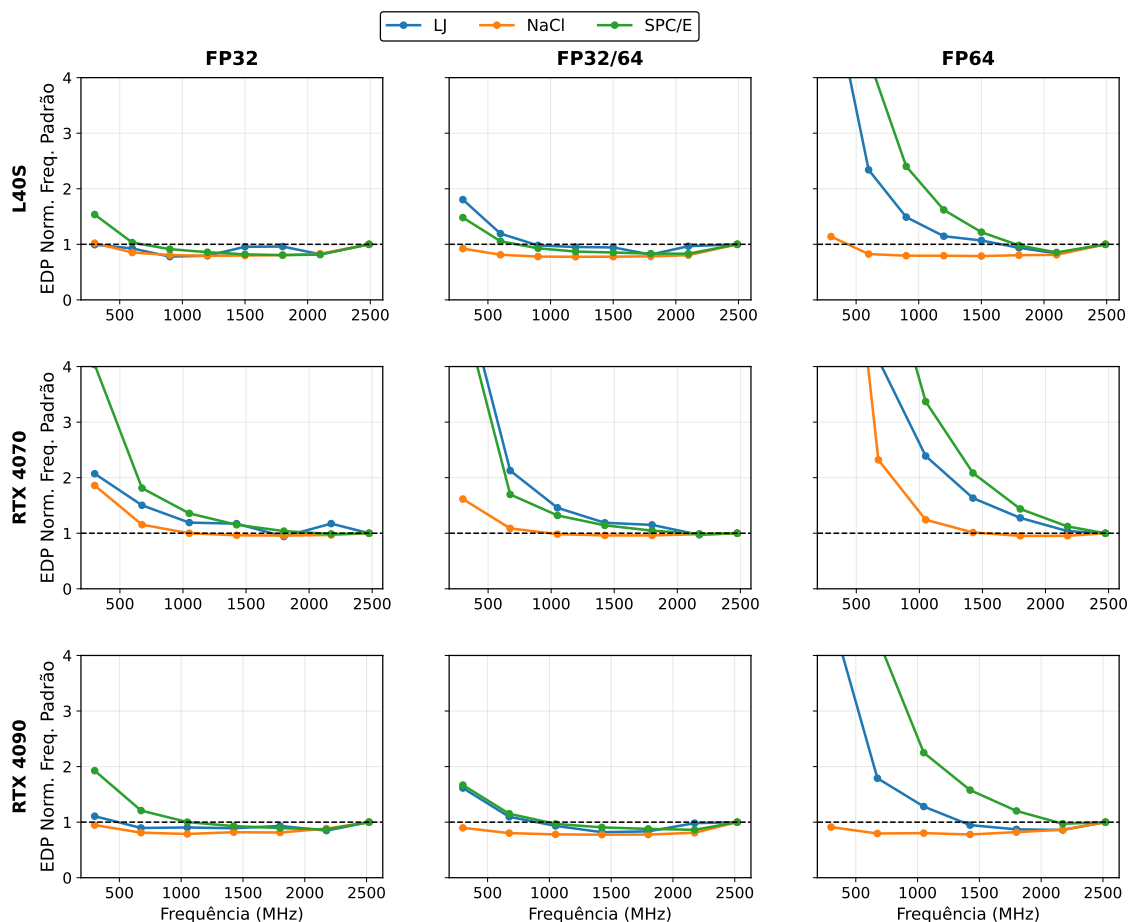


Figura 3. EDP normalizado ao valor observado na maior frequência avaliada para cada combinação de carga, precisão e GPU. Valores inferiores a 1 indicam que a redução de frequência melhora o EDP. O eixo vertical é limitado a $4\times$ para destacar a região energeticamente relevante.

os três modos numéricos, enquanto no SPC/E da RTX 4090 ela permanece em 2175 MHz. O LJ da RTX 4090 ainda apresenta um comportamento distinto: FP32/64 desloca o mínimo para 1425 MHz, enquanto FP32 retorna a 2175 MHz. Assim, a redução de precisão modifica especificamente o balanço entre desempenho e potência de cada combinação entre carga e arquitetura. A frequência de menor EDP difere da observada em FP64 em cinco das nove combinações, com redução mediana de aproximadamente 14% nos casos em que ocorre deslocamento. Portanto, mesmo alterações menos agressivas na representação numérica podem modificar a região de operação energeticamente eficiente.

Assim, precisão e DVFS não constituem mecanismos de otimização independentes. A escolha do modo numérico modifica a sensibilidade da aplicação à frequência e, consequentemente, o ponto em que a redução de potência compensa a perda de desempenho. Uma política de DVFS definida para FP64 não deve, portanto, ser reutilizada automaticamente após a adoção de FP32 ou FP32/64; o ponto energeticamente eficiente deve ser determinado considerando conjuntamente a carga, a precisão e a arquitetura.

Tabela 2. Redução de EDP obtida com FP32 e o melhor ajuste de DVFS mantendo FP64.

GPU	Carga	FP32	DVFS em FP64	Maior efeito
L40S	LJ	68,8%	16,3%	Precisão
	SPC/E	31,4%	14,6%	Precisão
	NaCl	8,1%	21,1%	DVFS
RTX 4070	LJ	88,6%	0,0%	Precisão
	SPC/E	84,2%	0,0%	Precisão
	NaCl	15,4%	4,7%	Precisão
RTX 4090	LJ	73,6%	14,0%	Precisão
	SPC/E	23,1%	2,9%	Precisão
	NaCl	10,6%	22,2%	DVFS

4.4. Impacto da Precisão vs DVFS no EDP

Os resultados anteriores mostram que precisão e frequência oferecem oportunidades distintas para a redução do EDP. Para avaliar a importância relativa desses mecanismos, a Tabela 2 compara o ganho obtido ao substituir FP64 por FP32 na configuração de referência, com a maior redução de EDP obtida ao manter FP64 e ajustar apenas a frequência. Os percentuais representam, em cada caso, a redução em relação ao respectivo *baseline* do experimento.

A comparação revela uma hierarquia de mecanismos diretamente relacionada ao comportamento da carga. No LJ, a redução de precisão é claramente dominante em todas as três GPUs. Na L40S e na RTX 4090, FP32 reduz o EDP em 68,8% e 73,6%, enquanto o melhor ajuste de frequência, mantendo FP64, alcança apenas 16,3% e 14,0%. Na RTX 4070, o contraste é ainda maior: o melhor ponto de FP64 já ocorre na maior frequência avaliada, enquanto FP32 reduz o EDP em 88,6%. O SPC/E segue predominantemente o mesmo padrão, embora com menor separação na L40S e na RTX 4090.

O NaCl é a principal exceção e reforça o mecanismo identificado anteriormente. Como essa carga praticamente não transforma o maior *throughput* de FP32 em menor tempo de execução, a redução de precisão oferece pouco espaço adicional de otimização. Nesse caso, DVFS torna-se o mecanismo mais efetivo na L40S e na RTX 4090, reduzindo o EDP em 21,1% e 22,2%, contra 8,1% e 10,6% com FP32. Portanto, não existe um mecanismo universalmente dominante. Isto é, quando a carga é sensível à capacidade computacional, a precisão constitui o primeiro parâmetro a ser explorado; quando essa sensibilidade desaparece, o ajuste de frequência pode produzir um retorno energético maior.

Essa distinção também mostra por que precisão e DVFS devem ser vistos como mecanismos complementares, e não concorrentes. Mesmo após a adoção de FP32, o ajuste de frequência ainda proporciona reduções adicionais de EDP que chegam a aproximadamente 22% em algumas configurações. Assim, uma estratégia eficiente deve, primeiro, identificar qual mecanismo apresenta maior potencial para o regime de aplicação e, posteriormente, explorar sua combinação com o segundo parâmetro.

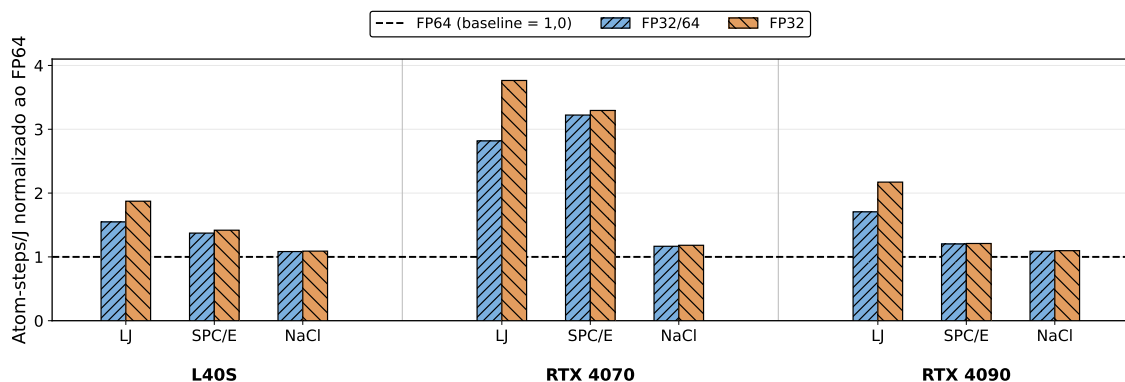


Figura 4. Eficiência científica em *atom-steps* por joule, normalizada ao FP64 da mesma carga e arquitetura. Valores superiores a 1 indicam maior quantidade de trabalho científico concluído por unidade de energia.

4.5. Eficiência científica por unidade de energia

O impacto prático dessas diferenças pode ser expresso em trabalho científico por joule, conforme definido na Seção 3. A Figura 4 mostra essa eficiência normalizada em relação ao FP64. No LJ, FP32 permite realizar $1,87\times$, $3,76\times$ e $2,17\times$ mais *atom-steps* por joule na L40S, na RTX 4070 e na RTX 4090, respectivamente. FP32/64 preserva parte substancial desse ganho, alcançando $1,55\times$, $2,82\times$ e $1,71\times$. Esses são justamente os casos em que a menor precisão reduz simultaneamente o tempo e a potência.

No NaCl, a eficiência cresce apenas para $1,09\times$, $1,18\times$ e $1,10\times$ em FP32. Como praticamente não há redução no tempo de solução, o ganho fica limitado à menor energia consumida. O SPC/E reforça a influência da plataforma: FP32 alcança $1,42\times$ na L40S e $1,21\times$ na RTX 4090, mas $3,29\times$ na RTX 4070. Nesta última, FP32/64 atinge praticamente o mesmo resultado, com $3,22\times$. Esses resultados mostram que reduzir potência, isoladamente, não implica produzir substancialmente mais ciência com a energia disponível. Os maiores ganhos ocorrem quando a menor precisão também reduz o tempo necessário para concluir a simulação, permitindo converter o mesmo orçamento energético em uma quantidade maior de trabalho científico.

Por fim, esse aumento de eficiência não ocorre às custas da validade das simulações. Os modos FP32 e FP32/64 utilizados neste estudo foram previamente validados pelos autores quanto às propriedades físicas e numéricas relevantes das cargas consideradas. Embora as diferenças de arredondamento sejam inerentes à alteração da representação numérica, elas não modificam os resultados científicos segundo os critérios de validação adotados. Assim, os ganhos de *atom-steps* por joule representam uma redução efetiva do custo energético necessário para produzir resultados científicos equivalentes.

5. Conclusão

A eficiência energética de aplicações científicas não deve ser avaliada apenas pela redução da potência instantânea, mas também pela quantidade de trabalho científico concluído com a energia disponível. Em dinâmica molecular, essa distinção é particularmente importante porque uma configuração de menor potência pode aumentar o tempo de execução e, conseqüentemente, reduzir pouco o custo energético total da simulação. Sob essa perspectiva, maximizar a ciência por joule fornece uma interpretação mais direta da

eficiência, em que o objetivo passa a ser concluir um número maior de simulações válidas com o mesmo orçamento energético, preservando simultaneamente desempenho e fidelidade numérica.

Este trabalho mostrou que a precisão numérica é um mecanismo importante para alcançar esse objetivo, mas seu benefício depende fortemente da carga e da arquitetura. Em FP32, a redução de EDP variou de apenas 8,1% a 88,6% entre as configurações avaliadas, resultado explicado pela capacidade, ou não, de converter o maior *throughput* da menor precisão em redução do tempo de execução. Esse efeito levou a ganhos de até $3,76\times$ em *atom-steps* por joule no LJ e $3,29\times$ no SPC/E, ambos na RTX 4070, enquanto o NaCl permaneceu próximo de $1,1\times$. Mostramos ainda que FP32/64 preserva grande parte desses benefícios e que a frequência de menor EDP se desloca quando a precisão é alterada. Portanto, precisão, características da carga e frequência de operação devem ser consideradas conjuntamente para maximizar a quantidade de ciência produzida por unidade de energia.

Como trabalhos futuros, pretendemos ampliar a caracterização para outras arquiteturas de GPU e sistemas de maior escala, avaliando se os mecanismos identificados se mantêm sob diferentes relações entre capacidade computacional, largura de banda de memória e desempenho em FP64. Também pretendemos expandir o conjunto de sistemas moleculares e investigar execuções mais longas, incorporando critérios adicionais de fidelidade física. Por fim, os resultados deste estudo abrem espaço para mecanismos automáticos de seleção conjunta de precisão e frequência, capazes de identificar, com base nas características da carga, a configuração que maximiza o trabalho científico realizado por joule.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código de Financiamento 001, FAPERGS – PqG 24/2551-0001388-1 e CNPq. Alguns experimentos deste trabalho utilizaram os recursos da infraestrutura PCAD, <http://pcad.inf.ufrgs.br>, no INF/UFRGS.

Referências

- Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., and Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2:19–25.
- Afzal, A., Hager, G., and Wellein, G. (2026a). Wattlytics: A web platform for co-optimizing performance, energy, and TCO in HPC clusters. Preprint, arXiv:2604.08182 [cs.DC].
- Afzal, A., Kahler, A., Hager, G., and Wellein, G. (2026b). GROMACS unplugged: How power capping and frequency shapes performance on GPUs. In *Euro-Par 2025: Parallel Processing Workshops*, volume 16294 of *Lecture Notes in Computer Science*, Cham. Springer. No prelo. Preprint: arXiv:2510.06902.
- Colberg, P. H. and Höfling, F. (2011). Highly accelerated simulations of glassy dynamics using GPUs: Caveats on limited floating-point precision. *Computer Physics Communications*, 182:1120–1129.

- Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, 103(19):8577–8593.
- Fox, J. and Weisberg, S. (2019). *An R Companion to Applied Regression*. SAGE Publications, Thousand Oaks, CA, 3 edition.
- Godoy, W. F., Hernandez, O., Kent, P. R. C., Patrou, M., Asifuzzaman, K., Miniskar, N. R., Valero-Lara, P., Vetter, J. S., Sinclair, M. D., Lowe-Power, J., and Bruce, B. R. (2026). Characterizing GPU energy usage in exascale-ready portable science applications. In *High Performance Computing. ISC High Performance 2025 International Workshops, Revised Selected Papers*, volume 16091 of *Lecture Notes in Computer Science*, pages 177–190, Cham. Springer.
- Hockney, R. W. and Eastwood, J. W. (1988). *Computer Simulation Using Particles*. Adam Hilger, Bristol.
- Hoefler, T. and Belli, R. (2015). Scientific benchmarking of parallel computing systems: twelve ways to tell the masses when reporting performance results. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '15, New York, NY, USA. Association for Computing Machinery.
- Johansson, A., Weinberg, E., Trott, C. R., McCarthy, M. J., and Moore, S. G. (2025). LAMMPS-Kokkos: Performance portable molecular dynamics across exascale architectures. In *Proceedings of the SC '25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1217–1232, New York, NY, USA. Association for Computing Machinery.
- Kashi, A., Lu, H., Brewer, W., Rogers, D., Matheson, M., Shankar, M., and Wang, F. (2026). Mixed-precision numerics in scientific applications: survey and perspectives. *The Journal of Supercomputing*, 82(5):287.
- Le Grand, S., Götz, A. W., and Walker, R. C. (2013). SPFP: Speed without compromise—a mixed precision model for GPU accelerated molecular dynamics simulations. *Computer Physics Communications*, 184(2):374–380.
- Machado, R. R. L., Eitzinger, J., Hager, G., and Wellein, G. (2026). On the challenges of energy-efficiency analysis in hpc systems: Evaluating synthetic benchmarks and gromacs. In *Proceedings of the Supercomputing Asia and International Conference on High Performance Computing in Asia Pacific Region Workshops*, SCA/HPCAsiaWS '26, page 50–58, New York, NY, USA. Association for Computing Machinery.
- Mei, X., Wang, Q., and Chu, X. (2017). A survey and measurement study of GPU DVFS on energy conservation. *Digital Communications and Networks*, 3(2):89–100.
- NVIDIA Corporation (2023). NVIDIA Ada GPU architecture. Whitepaper, v2.02. Disponível em: <https://images.nvidia.com/aem-dam/Solutions/geforce/ada/nvidia-ada-gpu-architecture.pdf>.
- Thompson, A. P., Aktulga, H. M., Berger, R., Bolintineanu, D. S., Brown, W. M., Crozier, P. S., in 't Veld, P. J., Kohlmeyer, A., Moore, S. G., Nguyen, T. D., Shan, R., Stevens, M. J., Tranchida, J., Trott, C., and Plimpton, S. J. (2022). LAMMPS—a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271:108171.

- Trott, C. R., Winterfeld, L., and Crozier, P. S. (2010). General-purpose molecular dynamics simulations on GPU-based clusters. arXiv:1009.4330.
- Verlet, L. (1967). Computer “experiments” on classical fluids. I. thermodynamical properties of Lennard-Jones molecules. *Physical Review*, 159(1):98–103.
- Yang, Z., Adámek, K., and Armour, W. (2024). Accurate and convenient energy measurements for gpus: A detailed study of nvidia gpu’s built-in power sensor. In *SC24: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–17.