Prezados membros da Comissão de Licitação

Mais uma vez, agradecemos toda a atenção e esforços para concluirmos a presente licitação e dotarmos o Laboratório Multiusuário de Computação Científica da UFLA de moderno recurso computacional. Nosso intuito é justamente garantir que teremos equipamentos adequados e que supram nossa demanda e necessidade de processamento, considerando os vários grupos de pesquisa da UFLA que trabalham com Computação Científica nas áreas de Química, Física, Biologia, Engenharia e Ciência da Computação.

Na presente oportunidade, fizemos uma análise comparativa das máquinas e os dois pontos que observamos grande diferença são i) processadores e a ii) GPU (GPU Node). Dessa forma, seguem as observações:

i) Processadores

i.1) Os processadores do *Head Node* possuem o mesmo número de cores, porém o clock é bem menor, **o que influencia bastante no desempenho**, pois, como é de amplo conhecimento, o gerenciamento do cluster é serial e por isso a configuração de referência no edital possui um clock alto e poucos cores. Dessa forma**, com relação esse item, não se observa equivalência com a configuração de referência presente no edital**.

i.2) Outra diferença que pode ser observada com relação a configuração de referência descrita no edital é sobre os processadores do *Compute Nodes* (Nós de processamento) que possuem 4 cores a menos por processador, porém, um clock mais alto.

i.3) Outra diferença que pode ser observada com relação a configuração de referência descrita no edital é que o GPU Node (GPU) possui 8 cores a mais por processador, porém, um clock mais baixo.

i.4) Em termos dos nós de processamento (*Compute Nodes)*, utilizamos, conforme comentamos e **consta no edital**, o benchmark HPL para os testes de desempenho. De fato, o spec é um ótimo parâmetro para avaliação de desempenho no geral, mas por se tratar de computação científica e computação de alto desempenho (HPC), estes benchmarks utilizados acabam sendo muito distintos dos workloads que executamos no laboratório. Assim, atualmente, o HPL é o parâmetro mais reconhecido de avaliação de desempenho de clusters de alto desempenho. **Os resultados de desempenho não foram apresentados com o benchmark HPL, que consta no edital.**

i.5) Cálculo aproximado de desempenho com precisão dupla, o pico real da solução proposta está descrito abaixo:

Tabela 1. Desempenho aproximado utilizando o pico real da solução proposta.

Node	Est. Rmax
Compute Nodes	3,89TFLOPS
GPU Node	0,47TFLOPS

Tabela 2. Desempenho aproximado utilizando o pico real da solução que consta no termo de referência.

Node	Est. Rmax
Compute Nodes	4,32TFLOPS
GPU Node	1,29TFLOPS

O cálculo de pico real foi feito baseado em experiências e em máquinas da mesma família de processadores para ambas soluções. Dados acima ainda estão sem os valores das GPUs, pois a Nvidia omitiu os valores da T4 para precisão dupla. É

importante ressaltar também que o *Head Node* não foi calculado, pois não fará processamento de dados. Dessa forma**, com relação esse item, não se observa equivalência com a configuração de referência presente no edital**.

i.6) Notamos também que o cache e a velocidade das memórias **são inferiores às especificações do edital.**

ii) GPU Node

ii.1) A parte da GPU é o ponto mais crítico, pois utilizaremos justamente para o software Gaussian, que é o software mais utilizado na UFLA para cálculos de estrutura eletrônica, e a T4 não está homologada para esse software, conforme documento do site oficial https://gaussian.com/g16/gpus.pdf. Além do mais, não encontramos nenhum artigo utilizando a mesma. Esse fato certamente compromete a instalação do Gaussian no cluster de computadores.

No edital, a instalação do Gaussian no cluster de computadores é solicitada no ítem 3.1 de Serviços Inclusos. A GPU é parte integrante do cluster, que consta de 1 *Head Node*, 2 *Computer Nodes* e 1 *GPU Node*. Nesse sentido, é importante salientar que, para garantir a instalação e o pleno funcionamento do Gaussian no cluster, a GPU (GPU node) deve ser homologada para o software Gaussian. A configuração descrita no termo de referência, que consta no edital, possui GPU (GPU node - P100) homologada para o software Gaussian. A instalação e o pleno funcionamento do software Gaussian como também a equivalência com a GPU de referência, que é prerrogativa no edital, foi solicitada para o pleno funcionamento do mesmo em nosso workload. **Dessa forma, com relação a esse item, não se observa equivalência com a configuração proposta como referência.**

ii.2) Além disso, no comparativo entre a GPU T4 e P100, os testes mostraram desempenho muito abaixo da T4 tanto para resultados de precisão simples quanto dupla. Provavelmente, por não ser da categoria HighEnd da Nvidia Tesla. Por isso, em termos de desempenho, ela acaba sendo muito inferior a P100, que consta no edital como referência. Os dados são descritos nas tabelas 3 e 4.

Tabela 3. Double Precision Results

GPU	Tesla T4	Tesla V100	Tesla P100
Max Flops (GFLOPS)	253.38	7072.86	4736.76
Fast Fourier Transform (GFLOPS)	132.60	1148.75	756.29
Matrix Multiplication (GFLOPS)	249.57	5920.01	4256.08
Molecular Dynamics (GFLOPS)	105.26	908.62	402.96
S3D (GFLOPS)	59.97	227.85	161.54

Tabela 4. Single Precision Results

GPU	Tesla T4	Tesla V100	Tesla
Max Flops (GFLOPS)	8073.26	14016.50	9322.46
Fast Fourier Transform (GFLOPS)	660.05	2301.32	1510.49
Matrix Multiplication (GFLOPS)	3290.94	13480.40	8793.33
Molecular Dynamics (GFLOPS)	572.91	997.61	480.02
S3D (GFLOPS)	99.42	434.78	295.20

Link completo: https://www.microway.com/hpc-tech-tips/nvidia-turing-tesla-t4-hpc-performance-benchmarks/

Dessa forma, com relação esse item, não se observa equivalência com a configuração proposta como referência como descrito no item 2.1.1. de Justificativa e Objeto da Contratação.