

VLAB@UFSC

Laboratório Virtual Multiusuário de
Ensino, Extensão e Pesquisa

Inteligência Artificial, Genômica,
Computação Quântica e
de Alto Desempenho



Por quê?

O desenvolvimento acelerado das técnicas e aplicações de Inteligência Artificial (IA), tanto na área de processamento de linguagem como nas áreas de visão computacional e geração de imagem tornaram imperativa a inserção de atividades práticas de uso e de desenvolvimento de ferramentas de IA em todos os cursos de graduação e pós-graduação, não apenas em cursos de STEM.

O mesmo diz respeito à Computação Quântica (CQ) e à Genômica Computacional (GC). Esta última é assunto com o qual o estudante de medicina, biologia, micologia, botânica, zoologia, agronomia e microbiologia tem de ser familiarizado já desde o início da sua formação.

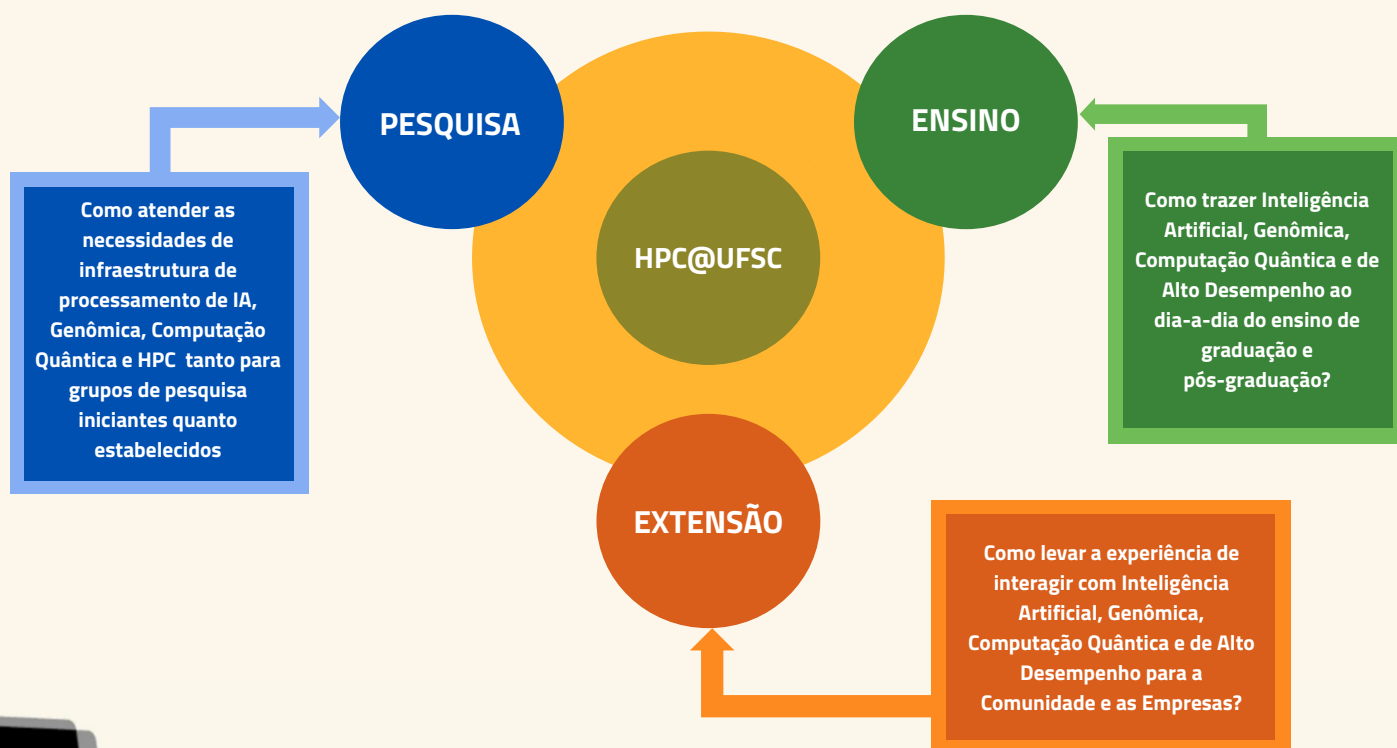
Ao mesmo tempo precisamos fomentar a disseminação fundamentada da IA, facilitando P&D&I e atividades de Extensão dentro da Universidade.

Como?

Oferecendo Infraestruturas Multiusuário de Ensino, Extensão e Pesquisa em IA, QC, Genômica e HPC para redes regionais de Universidades, de forma que:

- **P&D&I** seja suportado com recursos regionais, sem necessidade de aluguel de infraestrutura em nuvem, inacessível a grupos emergentes;
- **Ensino** com atividades práticas de laboratório seja viabilizado através de infraestrutura compartilhada em nuvem local e com software básico acessível ao aluno e ao professor;
- **Extensão** com atividades práticas hands-on possa ser realizada com a Comunidade e com a Indústria, disseminando novas tecnologias.

Datacenters regionais oferecerão esse serviço.



COMPUTAÇÃO
NA ESCOLA



LAPIX
Laboratório de Processamento
de Imagens e Computação Gráfica

INCoD
Instituto Nacional para Convergência Digital



UNIVERSIDADE FEDERAL
DE SANTA CATARINA



fapesc
Fundação de Amparo à
Pesquisa e Inovação do
Estado de Santa Catarina

Infraestrutura: Software Básico + Hardware

VLAB/UFSC - Laboratório Virtual para P&D&I, Ensino e Extensão

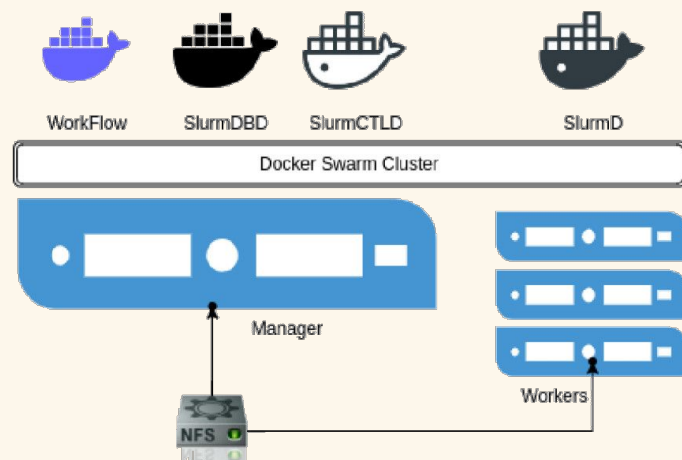
Movidos pelas demandas da área de pesquisa, a instituição desenvolveu um serviço de processamento científico baseado em containers, premiado e reconhecido pelo CGTIC, que sana as demandas dos grupos de pesquisa. Tal iniciativa, apesar de flexível, ainda não atende por completo as demandas de pesquisa, ensino e extensão devido a limitação de recursos de hardware que a instituição possui.

Idealizamos o serviço para, no ponto de vista do usuário, ser o mais simples e funcional possível, inspirados em soluções como Google Colab. A infraestrutura utiliza as tecnologias de software livre Jupyter Hub, Docker e SLURM. Investiu-se em uma solução que atendesse as demandas da comunidade universitária, assim desenvolvemos um auto-serviço para receber, enfileirar e executar as demandas de trabalho (Jobs) de acordo com sua ordem de chegada e políticas de prioridade definidas pelo serviço. A parte de processamento de dados ocorre em instâncias de Jupyter Notebooks e parte em clusters Docker Swarm contendo um manager e N workers, ilustrado na figura ao lado.

Hardware de HPC (GPGPU)

Para a realização deste projeto deverá ser adquirido um equipamento de HPC com arquitetura NVIDIA HGX/H100. Permite configuração flexível de oito placas H100 de 80 GB VRAM cada, possibilitando a realização de treinamento de modelos de IA e de simulações de CQ que necessitem de até 640 GB VRAM para a sua execução, permitindo também a compartimentalização via software, de maneira que vários grupos de usuário possam enxergar partes do equipamento como uma VRAM independente na qual podem executar as suas simulações e treinamentos. Pode tanto atender algum subprojeto que demande uma quantidade massiva de recursos, quanto atender ao mesmo tempo uma grande quantidade de projetos menores e de atividades de ensino e extensão, todas em paralelo. Uma visão geral da infraestrutura que desejamos adquirir pode ser obtida nesta brochura, <https://blog.nvidia.com.br/2022/06/21/apresentando-a-nvidia-hgx-h100/>.

VLAB será hospedada na Superintendência de Governança Eletrônica e Tecnologia da Informação e Comunicação (SeTIC) da UFSC, que gerencia os Datacenters da instituição, mantendo-os sob padrões nacionais e internacionais



Notebook para Segmentação Semântica baseada em Transformers (PyTorch)

Outubro de 2022

Autores:

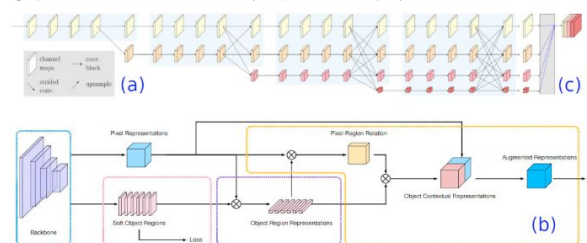
- Aldo von Wangenheim: aldo.vw@ufsc.br
- Luc Antonio Buschetto Macarini: luiz.buschetto@proceed.ufsc.br
- Thiago Zimmermann Loureiro Chaves: thiago.zc@proceed.ufsc.br
- Rodrigo de Paula e Silva Ribeiro: rodrigo.rps@proceed.ufsc.br
- Nathalie Ferreira: nathalie@proceed.ufsc.br

Notebook Multi-Transformer

HRNet

Apesar de tecnicamente as Redes Neurais de Alta Resolução (High Resolution Neural Network - HRNet) não serem um modelo de transformador visual, a sua variação mais recente, a HRNet+OCR (HRNet plus Object Context Recognition) possui um mecanismo de atenção primitivo a título de pós-processamento que faz com que vários autores as classifiquem na categoria de transformadores visuais (esse é o caso do pacote PaddleSeg, que a incorpora, chamando-a de OCRNet). HRNet é uma rede neural de arquitetura homogênea, linear e sequencial, no melhor estilo das redes neurais VGG, lá do começo da era do aprendizado profundo, mas que possui como diferencial o fato de dividir o fluxo de processamento em quatro fluxos paralelos de resoluções diferentes 1/1, 1/2, 1/4 e 1/8 de resolução. A linearidade da sequência de processamento é quebrada apenas por um conjunto de conexões verticais integrando os quatro fluxos em intervalos regulares.

A rede HRNet pode ser terminada em três tipos de cabeças diferentes, dependendo se a função é classificação de imagens, detecção de objetos ou segmentação semântica. A arquitetura para essas três categorias de aplicação se diferencia apenas pela cabeça ao final. É um modelo de rede neural que utiliza muita memória e normalmente são indicadas placas gráficas com 24 GB ou mais para treinar uma instância de HRNet, embora para conjunto de dados pequenos seja possível treiná-la em placas de 16 GB. De um ponto de vista de estética matemática, essa homogeneidade do modelo e estrutura simples e lógica que a HRNet utiliza a torna um modelo bonito e fácil de compreender, tornando o modelo adequado para fins didáticos também.



A figura acima ilustra as duas possíveis arquiteturas de HRNet: arquitetura clássica de 2019 (a) termina com um bloco que possui conjunto de operações de convolução 1x1 na posição (c) que integram os quatro fluxos de processamento multi-resolução e alimentam uma cabeça de classificação de imagens, detecção de objetos ou segmentação semântica. Já a versão mais moderna, HRNet+OCR, de 2021, possui na posição (c) um bloco Object Context



Pesquisa

IA, Genômica Computacional, Computação Quântica e HPC passaram a permear todas as áreas de pesquisa e não mais apenas STEM. P&D&I em IA, GC, CQ e HPC passou a ser essencial também para as Ciências Humanas, da Terra e da Vida.

Se a comunidade brasileira de P&D&I quiser ter um papel que vá além do de meros utilizadores e refinadores de tecnologias desenvolvidas por grupos do Exterior, é necessário que se ofereça recursos de processamento adequados e de acesso facilitado a toda esta comunidade, liberando-a da necessidade do aluguel de infraestrutura em nuvem no Exterior:

Grupos de pesquisa maiores, com projetos mais ambiciosos, dependem uma fração considerável de seus recursos no aluguel de infraestrutura em nuvem para o treinamento de modelos de IA ou a realização de simulações, com o risco de, havendo a necessidade de repetir uma simulação ou um treinamento, não restarem mais recursos no projeto para a realização dessa atividade;

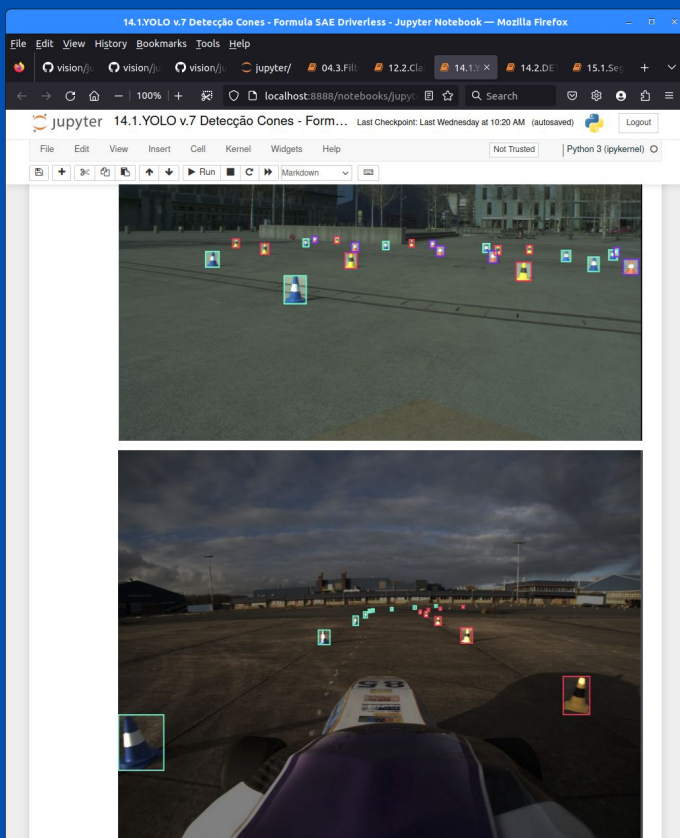
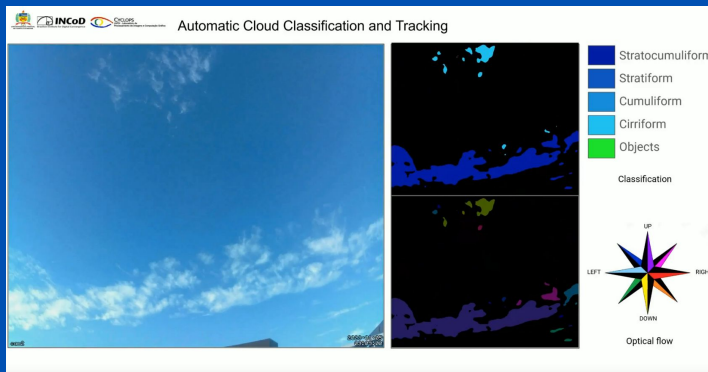
Grupos de pesquisa menores e pesquisadores iniciantes ou recém doutores sem recursos, muitas vezes não têm acesso à capacidade de processamento necessária para a realização de seus projetos em função do fato de que o que existe de oferta gratuita, como Google colab, não ser suficiente para a realização de seus treinamentos ou simulações ou possuir tempo de acesso limitado, insuficiente;

Grupos de pesquisa de áreas não-STEM são desestimulados pela complexidade, custos e dificuldade de acesso à infraestrutura, deixando de acompanhar o desenvolvimento internacional de sua área de pesquisa.

É necessária uma infraestrutura que ofereça:

Reserva de capacidade de processamento para atividades de pesquisa: VLAB deverá ser dimensionado de forma a contar com um conjunto de recursos físicos que possam ser reservados para pesquisa e seu software básico deverá possibilitar o gerenciamento e a reserva de segmentos de VRAM para uso por períodos mais prolongados por parte de projetos de pesquisa específicos;

Oferta de software aplicativo e de plataforma de desenvolvimento coerente com o estado da arte na pesquisa e no ensino em IA e HPC. Para que alunos de graduação e professores de áreas não-STEM possam utilizar a plataforma com conforto e facilidade, deverão ser oferecidas plataformas como Anaconda/Jupyter para desenvolvimento e experimentação e plataformas como Gradio para deploy de aplicações gráficas com GUI.



Ensino

A disseminação massiva da IA e outras técnicas de HPC em todas as áreas do conhecimento, com profunda influência sobre a prática profissional, tornou necessário colocar não apenas os alunos de STEM em contato com essas tecnologias, mas também alunos das áreas de Humanas e de Ciências da Terra e da Vida. Ter tido contato com HPC é hoje essencial para a formação profissional de todo estudante de graduação:

IA Gerativa tornou-se essencial nas áreas de Design, Animação, Games, Cinema e Artes, nas quais passou a ter um papel essencial na idealização e projeto de ambientes e personagens. O graduando de Animação e de Games necessita ser capaz de utilizar IAG como uma ferramenta na sua vida profissional e tem que ter tido a experiência prática de utilizá-la em sala de aula durante a graduação, tendo sido inclusive formalmente confrontado com todos os aspectos éticos relacionados ao uso dessa tecnologia;

Visão Computacional tornou-se essencial nas áreas da Agronomia, Engenharia Florestal, Agricultura de Precisão, Micologia, Botânica, Taxonomia Animal e Microbiologia. O aluno de um curso de Botânica, Engenharia Florestal ou Agronomia necessita ser capaz de utilizar ferramentas de VC para levantamento de cobertura vegetal ou cálculo de distribuição e biomassa de daninhas em agricultura de precisão;

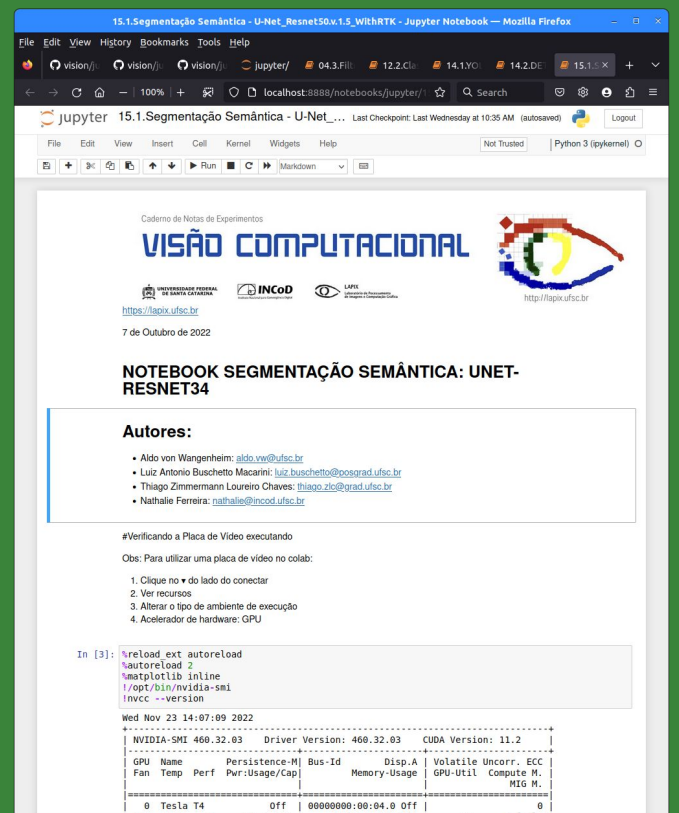
Podemos listar a necessidade da familiaridade com Simulação de Modelos Climáticos para o aluno de Oceanografia ou Meteorologia; familiaridade com ferramentas de Genômica Computacional para o aluno de Bioinformática, Medicina ou Técnica Forense; ou familiaridade com ferramentas de simulação de ligações moleculares com IA para o aluno de Farmacologia, Bioquímica ou Engenharia Química.

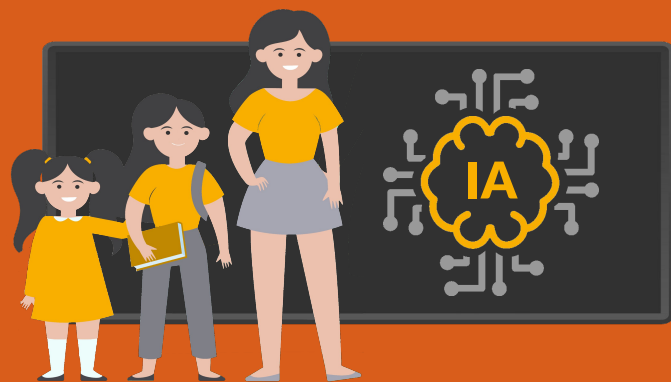
Para trazer IA, GC, CQ e HPC ao dia-a-dia do ensino de graduação e pós-graduação é necessária uma infraestrutura que facilite ao professor o preparo das aulas e seja amigável ao aluno. Necessitamos hardware tipo GPGPU flexível, compartimentalizável e capaz de suportar todas essas modalidades de software:

É necessário **Software Básico** que permita reservar, por fatias de tempo pré-definidas, parte das capacidades de HPC para diferentes turmas, de acordo com dimensões e necessidades de processamento de cada turma.

O software básico necessita possuir uma **interface de usuário simples e intuitiva**, permitindo também ao professor da área de Humanas ou de Ciências da Vida configurar o acesso dos alunos, testar o seu material de aula e planejar simulações e exercícios.

O software básico necessita também suportar a reserva de **capacidade de processamento fora de horários de aula** para os alunos finalizarem suas simulações e realizarem trabalhos.





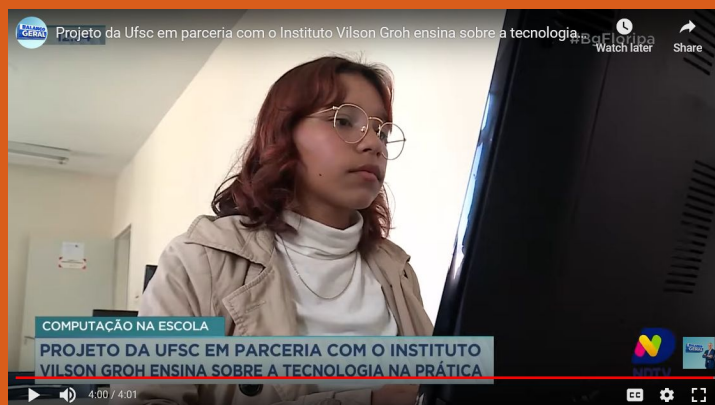
Extensão

Computação é uma das habilidades essenciais do século XXI. A publicidade, muitas vezes fantasiosa ou negativa, que a IA tem recebido na mídia nos últimos anos, tem criado mitos e ideias equivocadas e inclusive animosidade aberta por parte de segmentos da população menos informados. É função da Universidade esclarecer, informar e educar também a Comunidade, disseminando conhecimento científico e tecnológico, criando oportunidade de todos os segmentos da população de terem contato hands-on com as novas tecnologias.

Cursos de Extensão Universitária abertos que possibilitem o contato prático da Comunidade com as tecnologias de IA são uma forma muito importante de disseminação da C&T, de execução de atividades de Ciência Cidadã e de familiarização e capacitação focada de segmentos da sociedade considerados chave, como por exemplo professores e alunos do Ensino Público Fundamental e Médio e grupos especiais de estudantes em um contexto de vulnerabilidade social, que de outra forma jamais teriam um acesso hands-on a tecnologias de IA. Também áreas ainda mais avançadas como a Computação Quântica necessitam ser disseminadas e desmistificadas através da possibilidade de o público interessado ter contato com elas.

Por outro lado, a Extensão Universitária na forma de Projetos de Consultoria e de Transferência de Tecnologia é um fator chave na interação da Universidade com o setor produtivo e também no financiamento e na manutenção de grupos de pesquisa. A busca por parte do setor produtivo de consultorias nas áreas de IA e HPC intensificou-se nos últimos anos e é importante prover para as Universidades infraestrutura para que possam oferecer e executar projetos de consultoria de curta duração com agilidade, custos competitivos e sem a necessidade de aquisição ou aluguel prévios e demorados de hardware específico.

Para que a universidade possa ter essa interação ágil tanto com o setor produtivo quanto com membros da Comunidade, é necessário que exista uma infraestrutura presente regionalmente e que possa ser alocada, atendendo as necessidades dos grupos de pesquisa envolvidos em atividades de Extensão.



Disponibilizar uma estratégia como a que propomos para redes regionais de universidades vai não só alavancar massivamente as pesquisas de ponta utilizando HPC em Inteligência Artificial e outras áreas que usam IA mas possuem pouco acesso como as Humanas e Ciências da Vida, mas também transformar profundamente o Ensino e a Extensão, realizando a mais que necessária aproximação de alunos e Comunidade à experiência prática da utilização dessas tecnologias de ponta.

Capacidade permanente para Ensino & Extensão							VRAM (GB):	640
Requisitos e usos expressos em blocos de VRAM contínuos (servidor: HGX/H100 - 8 x 80GB)								
Modulo (bloco VRAM)	GB VRAM/ grupo	alunos/ grupo	alunos/ turma	turmas (GR/PG)	Blocos VRAM	GB VRAM/ turma	Ensino/Extensão: Área de aplicação e disciplinas	
Basico	10	2	48	1	24	240	IA básica/CG intermediária/proc.paralelo e distribuído	
Intermediário	20	3	36	1	12	240	Visão Computacional (VC) básica/CG avançada (custom. SD)	
Avançado	40	4	24	1	6	240	IA avançada (LLMs)/VC avançada (ViTs)	
Ultra	80	5	15	1	3	240	IA ultra-avançada (LLMs, Multimodais e ViTs customizados)	

Capacidade permanente para Atividades de P&D&I				
Modulo (bloco VRAM)	GB VRAM/ grupo pesquisa	Grupos de pesquisa/Blocos VRAM p/pesquisa	VRAM dedicada p/pesquisa	Pesquisa: Área de aplicação e tipos de projetos
Basico	10	40	400	Processamento de imagens em pesquisa aplicada, genômica, seqüenciamento
Intermediário	20	20	400	Processamento de imagens em pesquisa aplicada, arte computacional com IA para Cinema, Animação e Computação Gráfica
Avançado	40	10	400	P&D&I de modelos de visão computacional com transformers, pesquisas em Linguística com uso de LLMs
Ultra	80	5	400	P&D&I de modelos de LLMs, multimodais LLM/Visão e modelos de VC com ViT

Como oferecer ao Ensino e à Extensão?

Com o objetivo de ilustrar de forma mais concreta os perfis de uso, tanto de P&D&I, quanto de Ensino e Extensão, e criar um conjunto de valores para ser utilizado numa série de simulações, elaboramos a tabela acima que ilustra quatro perfis de uso, tanto para Pesquisa, como para Ensino e Extensão, que consideramos típicos e deverão ser esperados com o início da oferta do serviço. Com a flexibilidade da plataforma de hardware e da infraestrutura de software básico desenvolvido pelo SeTIC/UFSC podemos atender também outros perfis que venham a ser necessários.

Realizamos simulações de alocação do recurso de processamento para um conjunto de disciplinas de graduação e pós-graduação com diferentes graus de avanço e aprofundamento. A tabela ilustra a aplicação mais típica e mostra que com a estratégia de distribuição e alocação de recursos desenvolvida pelo SeTIC/UFSC nós podemos:

Atender as necessidades de ensino prático com IA e HPC de mais de 900 alunos/semana, em disciplinas básicas, médias e avançadas, com 240 GB VRAM permanentemente reservados para essas atividades;

Concomitantemente atender de 40 grupos de pesquisa iniciantes, com P&D&I em Visão Computacional e IA Gerativa, até cinco grupos de pesquisa avançados com pesquisas em LLMs e Modelos Multimodais, com a possibilidade de alocação dinâmica e variável e sem prejuízo às atividades de ensino e extensão, com 400 GB VRAM permanentemente reservados a P&D&I

A infraestrutura VLAB é flexível e permite atividades especiais, como a reserva dos 640 GB, por uma fatia de tempo, para simulações de CQ de grande porte ou o treinamento completo de um extenso modelo Multimodal LLM/Imagem.

Horários (blocos de 2 horas)	GR básico	GR intermed.	GR intermed.	PG intermed.	PG avançado	Disciplinas & alunos
matutino						
8:00	GR B1	GR B1	GR I1	PG I1	PG A1	
10:00	GR B2	GR B2	GR I2			
almoço						
12:00	TGRB	TGRB	TGRI	TPGI	TPGA	
vespertino						
14:00	GR B3	GR B3	GR I3	PG I2	PG A2	
16:00	GR B4	GR B4	GR I4			
noturno						
18:00	GR B5	GR B5	GR I5	PG I3	PG A3	
20:00	GR B6	GR B6	GR I6			
antelucano						
22:00	TGRB	TGRB	TGRI	TPGI	TPGA	
0:00	TGRB	TGRB	TGRI	TPGI	TPGA	
2:00	TGRB	TGRB	TGRI	TPGI	TPGA	
4:00	TGRB	TGRB	TGRI	TPGI	TPGA	
6:00	TGRB	TGRB	TGRI	TPGI	TPGA	
Modalidade	graduação (GR)			pós-graduação (PG)		
Disciplinas atendidas	6	6	6	3	3	24
Alunos atendidos	288	288	216	108	72	972