Request for Technology Fee Funds: FY21

NOTE: A separate request should be made for each initiative.

I. Department Number/Department Name: 360 College of Computing

Title of Request *(please be brief):* CHIPC GPU Cluster Upgrade Initiative

Amount of Request *(formula from detailed budget below):* $111,523

Type of Proposal: Atlanta or Dist Lrng/Non-Atl Atlanta

Is this request similar to one funded in FY19 or FY20? No (Yes or No)

Are there installation/renovation costs associated with this request? No (Yes or No)

If "Yes" then indicate the source of approved funding:
(Note: Tech Fees are not allowed for installation/renovation)

Executive Summary of Request (100 words or less):

This request will upgrade existing Center for High-Performance Computing (CHIPC) GPU computing infrastructure to support a common high-performance computing infrastructure. We propose to acquire multiple next-generation AMD servers and a prototype Intel GPU server to support benchmarking and machine learning efforts.

Specific class and/or lab initiative(s) if applicable:

Contact person for this request (incl. phone #): Jeffrey Young (404-385-1513)

Responsible faculty for this request (incl. phone #) Edmond Chow, Alexey Tumanov, Richard Vuduc, Will Powell

Indicate priority per department if applicable: Number ____ of ____

Indicate priority per college or unit: Number 3 of 9

II. Impact on Students - Provide course title, course number, and anticipated enrollments:

Titles/Numbers of Course(s) CSE 4220/6220; ECE 2601/3601/4601 (VIP); CS 4803/7643, CS

Anticipated Enrollments

Graduate:	*100*	*(per* year)	sem or yr
Undergraduate:	100	(per year)	sem or yr
Total:	200		

The estimated percent use of the resources in the item by:

Students	90%
Faculty	5%
Other	5%
Total:	100%

Brief explanation of how estimate was achieved.

Estimated percentage figures are based on making the GPU and cluster resources available primarily for students during the Spring and Fall semester. We anticipate using the Slurm scheduler to prioritize student usage.

NOTE: Other impacts on students should be described in narrative to include benefits to the students affected.

III. Detailed Budget - Requested Items by Category *List separately any equipment, software, and other allowable expenses (see Tech Fee Guidelines). There is a formula in the "total column" that multiplies the number of items times the unit price. You may enter a figure into the total column if the unit pricing is not applicable. If you need additional rows, contact the Budget Office to receive a modified form. Software or data license proposals should indicate how many years the item has been funded through student tech fees in narrative.*

Supporting documentation is required *- Include price justification in some form, such as quotations, published price lists, etc. as a separate PDF attachment. All supporting information should be in a single PDF.*

	Proposed Number of Items	Estimated Price per Unit	Total ($)
AMD Server with GPUs	2	$31,001	$62,001
4U Host Intel Cascade Lake Server Capable of hosting 4 GPUs	2	$12,775	$25,550
Nvidia V100	4	$5,993	$23,972
			$0
			$0
Total *(linked to the total amount of request line above)*			$111,523

Please return form via e-mail in Excel format to: techfees@business.gatech.edu. Supporting information only in a PDF file.

IV. Narrative - Provide narrative justification for your intended use of the technology fee funds. Include narrative on how the education or research of the students will be enhanced. To include curricular, co-curricular, and extracurricular benefits expected to accrue to students through provision of this resource, including students outside the unit. Briefly state how information regarding similar technology use elsewhere on campus to benefit from lessons learned, to standardize, or differentiate, and to avoid duplication. Also include how the request aligns with the Strategic Plan of Georgia Tech.

We propose to acquire next-generation AMD and Intel GPU resources along with three host servers to update our high-performance computing and machine learning cluster support under the Center for High-Performance Computing (CHIPC). This cluster will initially be used by Atlanta-based students in multiple courses but can eventually be extended to support OMSCS students via remote access and cluster-based scheduling. We anticipate supporting undergraduate students and graduate students for coursework as well as making the resource available to all interested students via our existing TSO-supported testbed. This effort will provide easier access to novel hardware for students as well as reduce the need for individual professors to acquire and maintain this type of hardware for their classes.

GPUs have been deployed as part of Georgia Tech research clusters (such as CoC-IC's SkyNet) and the instructional CoC PACE-ICE cluster, but this request differs because it focuses on next-generation AMD and Intel GPUs. Current systems focus exclusively on NVIDIA GPUs like the Titan XP and Volta GV100, which provides a good basis for machine learning training and inference and high-performance computing. However, existing Georgia Tech resources do not reflect the growing diversity of GPU computing as evidenced by Intel's announcement of its Xe GPU line at the Supercomputing conference in 2019 and AMD's new line of Instinct GPUs and Epyc CPUs which will be deployed in the first "exascale-class" supercomputer, Frontier in 2022. Recent student requests to TSO for AMD systems and specifically AMD CPU and GPU systems have indicated that this diversity of GPU resources is important for their learning environment in the College of Computing and is increasingly relevant for full-time employment in these fields outside of Georgia Tech. This type of resource is also not easily supplied by existing cloud services which focus on Intel CPUs and NVIDIA GPUs almost exclusively.

To promote a diverse environment for students to investigate HPC acceleration of applications, data analytics, and machine learning, we propose to acquire two different types of GPU accelerators and host servers to host remote access for the requested devices. We propose to acquire AMD MI50 Instinct high-end GPU cards, AMD Epyc host servers, and Intel Xe GPUs and one Intel host server. These resources will be used by the co-PIs of this request to support classes ranging from ECE (2,3,4)601/(2,3,4)602/4603 - Vertical Integrated Project (VIP) classes run by Dr. Vuduc and Mr. Powell (Undergraduate Student Cluster Competition), CSE 6220/CX 4220: Introduction to High-Performance Computing and CSE 6230: HPC Tools and Applications run by Drs. Chow and Vuduc, and CS 8803: Systems for Machine Learning run by Dr. Tumanov. The new Systems for Machine Learning course specifically will look at the hardware aspects that make these architectures different and how a common runtime, scheduling, and programming environment can potentially be used to bridge hardware and software differences across diverse GPU resources.

This mix of systems is proposed to allow for students to study and use diverse GPU acceleration options in terms of high-performance computing and also to use for training and inference for special projects in machine learning-related classes like CS8803 (Systems for ML), CS 4803 and CS 7643 (Deep Learning) via our scheduling interface and hosted server. Tensorflow and PyTorch backends for each of these types of GPUs will be supported through Intel Optimization for Tensorflow and AMD's RocM stack for machine learning. All of these GPUs will be available for VIP students in the student cluster competition class and our high-performance parallel computing courses, CX 4220/CSE 6230, where they will be made available for students to implement and evaluate high-performance algorithms written in languages like HIP, SyCL and OpenMP.

We will host the proposed infrastructure as part of the TSO-managed CHIPC testbed. This testbed currently has 20-30 users and supports student researchers in addition to faculty and graduate students for a variety of hardware, including several NVIDIA GPU boxes. The CHIPC infrastructure will allow for scheduling of GPU resources and hosting of common ML and HPC software in a common network-accessible location. These two capabilities will remove one of the key barriers for student usage of these devices – typically student projects are limited by the need to acquire new accounts on a machine that may sit in a professor's lab. This typically requires both physical access to a lab on GT's campus as well as a custom login to be created for each new student. The CHIPC testbed uses LDAP-based user authentication to give students backed up storage, access to relevant software, and access to a common set of programming and debugging tools. A robust set of wiki pages are currently available for this infrastructure (https://github.gatech.edu/chipc/chipc-docs/wiki), and we provide a discussion group for students to request assistance.

Please return form via e-mail in Excel format to: techfees@business.gatech.edu. Supporting information only in a PDF file.