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):* Datacenter-grade Infrastucture for Machine Learning

Amount of Request *(formula from detailed budget below):* $248,682

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 proposal outlines a purpose-built resource for high level Machine Learning courses that need greater control of the environment. We believe that this resource will be complementary of some of the share resources already provided by the College and will be able to be reserved for dedicated use.

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

Contact person for this request (incl. phone #): Dan Forsyth (4-9014)

Responsible faculty for this request (incl. phone #) Alexey Tumanov, Ada Gavrilovska, Mustafa Ammar

Indicate priority per department if applicable: Number ____ of ____

Indicate priority per college or unit: Number 4 of 9

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

Titles/Numbers of Course(s) CS 8803 Special Topics, CS6250, CS7210

Anticipated Enrollments

Graduate:	*468*	*(per* yr) sem or yr
Undergraduate:	0	(per yr) sem or yr
Total:	468	

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.

The majority of usage will be students conducting classwork on this resource.

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 ($)
GPU Server	4	$52,273	$209,091
High Speed Switch	1	$9,555	$9,555
Server Acceleration Card	4	$7,509	$30,036
			$0
			$0
Total *(linked to the total amount of request line above)*			$248,682

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.

Modern computing is rapidly changing in the face of major turmoil on both technological and application fronts. Applications are becoming more complex, often relying on Machine Learning components, which come with tremendous and growing computational requirements. Combined with the end of Dennard scaling and slowdown of Moore's law, keeping up with such growing application demands requires extreme scaling out and hardware specialization. The net result is that all high-performance systems will soon comprise a mix of heterogeneous compute technologies interconnected via high-performance networks. The fact that this architectural trend is already ubiquitous even in common datacenter/cluster deployments, and not just niche HPC markets, is indicative of an ongoing paradigm shift, as the datacenter market has conventionally exclusively relied on few standardized high-volume hardware components.

A paradigm shift of that degree has disruptive effects on computing. It is now essential for application developers to effectively handle distributed computation and hardware heterogeneity, while infrastructure providers need to face the challenging reality of managing an increasingly more heterogeneous pool of resources. Our goal as educators is to establish a mechanism to produce the next generation of students equipped with a skillset that resonates with these new realities. We propose to provision a small-scale datacenter system prototype that emulates the datacenter-grade heterogeneous servers comprising GPUs and FPGAs for computationally intensive and latency-sensitive tasks, such as ubiquitously deployed ML training and inference kernels, as well as high-end networking gear necessary for low-latency communication. Such a deployment would offer students hands-on experience with a scaled-down representative of modern datacenter-grade environments.

State of the Art at Georgia Tech

Georgia Tech has clusters of machines with GPUs, managed by a batch scheduler (SLURM). We argue that this infrastructure is complementary, but insufficient for our instructional purposes. We break down the reasons:

- Failure isolation: as students implement their classroom projects, they will be developing low-level systems infrastructure itself, NOT the applications on top of an existing infrastructure. With low-level systems infrastructure projects (e.g., deploying a new ML inference accelerator on an FPGA), the common mode of operation involves high probability of intermittent failures that could easily disconnect the node or a subset of nodes from the main operating cluster, which is undesirable in a shared environment.
- Ambient system noise: while batch-managed time-shared environment suits the needs of non-interactive, HPC-style workloads (including ML training), latency sensitive applications are affected by the ambient system noise and require performance isolation for reproducible and informative project experimentation. ML inference with multiple models or datacenter-type data serving workloads are highly sensitive to interference from co-located workloads, affecting the key performance and quality metric of tail latency (i.e., 95th/99th percentile) by an order of magnitude. Systems projects that target autoscaling algorithms for such ML inference pipelines base their decisions on the accuracy of these inference profiles.
- Privileged access and bleeding edge software: for more advanced, graduate-focused courses, a range of educational projects will require privileged (root) access to the hardware to modify device drivers, install specialized software, or even a specialized OS (e.g., dataplane OS latency-optimized networking).

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