UC San Diego Data Science/Machine Learning Platform (DSMLP)

Empowering Students with Future-Forward Skills

Larry L. Sautter Award Proposal 2019

submitted by: **Valerie E. Polichar, Ph.D.**
Director of Academic Technology Services
Information Technology Services, University of California San Diego
858-534-1859 | vpolichar@ucsd.edu

Project lead: Adam Tilghman
Architect/Analyst, Academic Technology Services, Information Technology Services, University of California San Diego

Figure 1 – "Intro to Autonomous Vehicles" student Cai (Joo Min) Yeo; vehicle training enabled by DSMLP

IT Services Implementation Team:

- **Adam Tilghman**, Academic Technology Services, IT Services – Architect/Analyst, Academic Technology Services – *Project lead; design, build, implementation, training, outreach*
- **Paul Jamason**, Manager, Ed Tech Ecosystem Services – *Faculty technical consultation*
- **David Anderson**, Applications Programmer, Ed Tech Ecosystem Services – *Backend deployment, middleware development*
- **Maria Andrade**, Applications Programmer, Ed Tech Ecosystem Services – *UI development*
- **Miguel Rodriguez**, Ed Tech Analyst, Educational Technology Services – *Faculty outreach*

IT Services Executive Level Support:

- **Vince Kellen**, Ph.D., Chief Information Officer
- **Valerie E. Polichar**, Ph.D., Director of Academic Technology Services
- **Dan Suchy**, Director, Educational Technology Services

Campus Partners:

- **Larry Smarr**, Founding Director of the California Institute for Telecommunications and Information Technology (Calit2); Harry E. Gruber professor, Computer Science and Engineering
- **Thomas DeFanti**, Research Scientist, California Institute for Telecommunications and Information Technology (Calit2)
- **John Graham**, Senior Development Engineer, California Institute for Telecommunications and Information Technology (Calit2)
- **Jacobs School of Engineering**
- **Department of Cognitive Science**

The **Data Science/Machine Learning Platform (DSMLP)** meets the growing need by graduate and undergraduate students for high-end computation resources with an affordable CPU/GPU cluster for coursework, formal independent study, student research and special projects.

Web-based Jupyter Notebooks can be customized with course-specific modules and libraries such as SciPy, NumPy, PyTorch, TensorFlow, Keras, NLTK, and AllenNLP. More complex computation can be conducted via terminal/ssh logins, background batch jobs, or custom Docker containers; a full Linux/Ubuntu CUDA dev suite is available.

The challenge: Students need compute power!

Increasingly, undergraduate and graduate courses in many disciplines require students to engage in advanced computation. Be it for a Sociology class conducting text analysis, a Music class performing acoustic modeling or an Engineering class developing sophisticated drones, the need for GPUs is burgeoning. Data science is a growing discipline, and machine learning is now useful to most disciplines in some form.

But GPUs are expensive. A comparison by UC San Diego researcher Thomas DeFanti indicates that high-end double-precision GPUs can cost as much as $10K each, with a fitted-out 8-GPU machine (with an expected lifespan of 4-6 years) exceeding $400K. In the cloud, four years of an 8-GPU equivalent could cost even more; use of spot pricing to mitigate costs still yields a per-machine cost of $80-100K. This puts conventional GPU computing out of reach for most coursework. In early 2017, nursing along an aging and modest student cluster, we were grappling with this fast-increasing need and looking for effective ways to meet it.

The solution:
Innovative research partners with production IT services to support instruction

Research work by UCSD faculty Larry Smarr and Thomas DeFanti on the Pacific Research Platform has been exploring cluster designs that minimize cost through utilization of game-oriented single-precision GPU hardware. In 2017, Smarr and DeFanti came to IT Services with a novel proposition: that, in partnership with the School of Engineering, we develop a student instructional cluster using their basic FIONA8 (Flash I/O Network Appliance - 8 GPU) design and use it to meet the growing needs of Engineering (and other disciplines) students.

CIO Vince Kellen agreed, and committed five years of funding in support of the project. Jacobs School of Engineering and the Department of Cognitive Science provided additional funding to build the initial cluster; equipment orders were placed in May of 2017. The stated goal was to meet the needs of graduate student classes, with undergrad classes supported as cycles were available. As the cluster came online in autumn of 2017, however, the demand for support for undergraduate coursework was immediate.

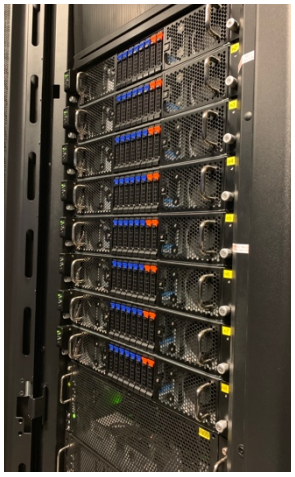

Figure 2 - The DSMLP

We worked closely with DeFanti's lab to adapt the FIONA8 design for the needs of large courses, and took advantage of IT Services' existing skill sets in system deployment and administration, user provisioning, and storage management. An initial $150k outlay was split evenly between IT Services and the academic units, allowing for 80+ GPUs, dataset storage, and networking. Since then, $50k/year of campus core funds have supported maintenance and modest year-on-year capacity growth.

On the software side, we provide multiple options for faculty, which makes the offering relevant across many disciplines that are not traditionally computation-focused. (This innovation has allowed our platform to appeal to a broad array of fields, as can be seen in Fig. 4 below.) Faculty can present students with their choice of stock or customized Jupyter Notebooks (most common among

undergraduate courses), terminal logins (popular for graduate courses and projects), or student-supplied Docker containers as best fits their course requirements.

The impact

We planned to initially focus the platform on graduate classes in the Engineering and Cognitive Science disciplines that sparked its creation. One or two classes would be supported the first quarter; we would add more as we gained confidence and as faculty heard about the service. We made the decision to hold off on actually advertising the platform until we had a full year of pilot service under our belt.

Figure 3 – Undergraduate Jessica Yang and her autonomous vehicle

But the unmet need was greater than we imagined, and word quickly spread. We supported 12 grad and undergrad classes in our opening quarter, across eight departments, including some, such as Music, that we hadn't anticipated.

By the end of year one (2017-18), the project, theoretically still in a 'pilot' phase, had attracted instructors in a dozen disciplines and was clearly already in production mode, surpassing our expectations for success. Since our official 'production' deployment in Fall 2018, we have been gradually reaching out to new departments to alert them about the availability of the platform. With each outreach, new classes join in.

A sampling of DSMLP courses for the Spring 2019 term include:

- Introduction to Autonomous Vehicles
- Spatial Data Science and Applications
- Machine Learning for the Arts
- Neurobiology & Bioinformatics Laboratory
- Machine Learning for Physical Applications
- Introduction to Python
- Data Science in Practice
- Machine Learning for Image Processing
- Image and Video Restoration
- Senior Seminar - Visual Arts

To date, the Data Science/Machine Learning Platform has been used in 97 classes across 16 departments in five divisions at UC San Diego.

The UC San Diego DSMLP was selected as a 2019 EDUCAUSE Horizon Report exemplar for its impact on the student experience.

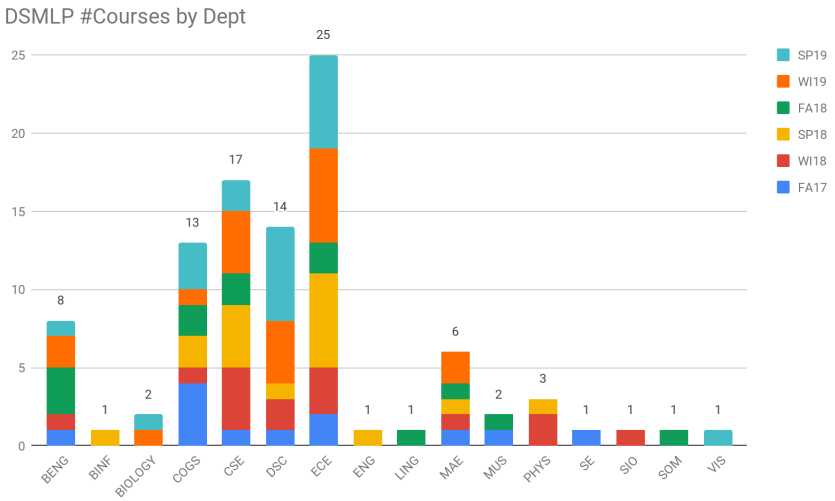

Figure 4 – Number of courses per department

Nearly 10,000 students have taken classes taught on the DSMLP, including 3000 in the current Spring 2019 quarter. In addition, we've accommodated undergraduate independent research, graduate thesis and dissertation research and student computing club computation needs.

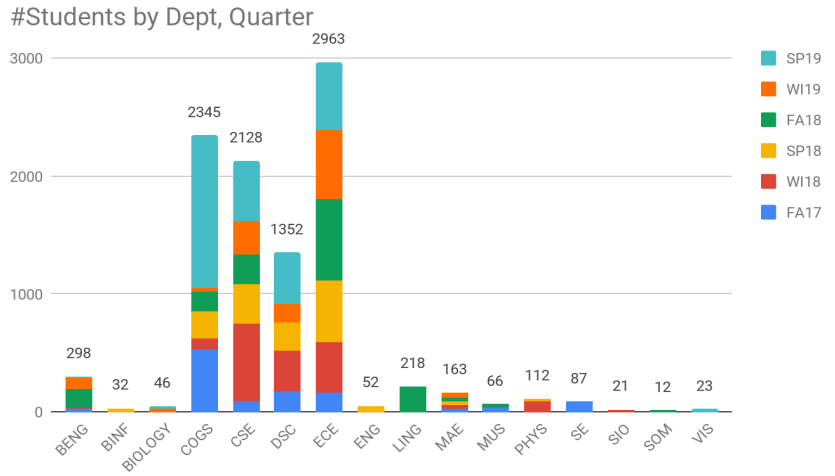

Figure 5 – Number of students per dept, quarter

The uptake speed for DSMLP has exceeded our expectations both in number of courses and students and in breadth of disciplines utilizing the platform. It provides unprecedented flexibility to faculty who may wish to have their students employ Web-based notebooks — or master power-user shell interfaces to direct computation. As we roll it out to further disciplines and courses, we envision wrapping most courses using our older time-share Linux platforms into the DSMLP, and expect we may eventually see as many as half of UC San Diego students making use of the platform at some point in their academic experience here.

Beyond the walls of UC San Diego

UC San Diego isn't keeping this innovation to ourselves. We make our technical design, budgets, and other details available to other UC campuses for the good of all UC students. Recently, we shared our experiences with UC Riverside, providing information that may help their IT staff build a solution appropriate for Riverside's own graduate and undergraduate needs.

References:

DSMLP website at UC San Diego

EDUCAUSE 2019 Horizon Report