Facilities, Equipment and Other Resources
Data Science Institute / Computer Science Department
Columbia University

The proposed effort will benefit from multifunctional shared spaces as well as state-of-the-art equipment, instrumentation, and computation and laboratory facilities that exist within the Data Science Institute, and other venues, like conference rooms and classrooms within Columbia University, that can be made available for the research, education and outreach initiatives outlined in the proposal. Note that all faculty researchers have access to work computers and office space.

The Data Science Institute is housed within 44,000 square feet of multifunctional space at Columbia University. This space includes faculty offices, conference rooms, student work stations and collaborative meeting spaces, all designed to promote a collegial and interactive physical community around data science and to encourage interdisciplinary approaches to research training. The Institute also has support staff to assist with the financial management of the project, as well as the educational and outreach activities associated with the project.

The Data Science Institute has access to Yeti, a High Performance Computing (HPC) Cluster housed in Columbia's centrally-managed Shared Research Computing Facility (SRCF), which consists of a dedicated portion of the university data center. The total server count is 167 execute servers with a total of 2762 compute cores. The CPU on all expansion machines is the 2.6 GHz E5-2650v2.

The Data Science Institute has also bought in two secure dedicated resources on the new Habañero cluster, which came online in November 2016. These resources consist of two "high memory nodes" and two "GPU nodes" along with 4TB of storage - whereas the overall specifications of the cluster are:

222 nodes with a total of 5328 cores (24 cores per node):

- 208 HP ProLiant XL170r Gen9 nodes w/dual Intel E5-2650v4 CPUs (2.2 GHz):
- 176 standard memory nodes (128 GB)
- 32 high memory nodes (512 GB)
- 14 HP DL380 Gen9 nodes with dual Intel E5-2650v4 Processors (2.2 GHz) and NVIDIA K80 GPU (2 per node) supplying ~140,000 GPU cores
- 407 TB DDN GS7K GPFS storage
- EDR Infiniband (FDR to storage)
- Red Hat Enterprise Linux 7
- Slurm job scheduler

Additionally, the Institute acquired from Yahoo! 40 HP compute servers to be used to support Institute research projects. These machines are configured with HP DL160 G5 2 x L5420 CPUs, 16GB of RAM and 1 x 1 TB 7.2K SATA Disk.

Shared storage is available through a Network Appliance storage file server.

The Computer Science departmental computing infrastructure consists of a data center for computational, storage and VoIP services, one large student laboratory, featuring 18 Windows machines and 33 Linux towers each with 2 Nehalem processors, 8 cores and 24GB memory; a remote Linux cluster with 17 servers each with 2 Xeon processors and 4GB memory, and a number of computing facilities for individual research labs. The data center houses 14TB of Network Appliance file servers, a compute cluster consisting of a Linux cloud with 32 servers each with 2 Nehalem processors, 8 cores and 24GB memory. This cloud can support approximately 5000 of VMware instances. Additional computing facility includes 8 Intel Linux servers with 8 processors and 16GB memory each and 3 Sun Fire 280R servers, a 3Com VoIP system, and departmental load balanced web cluster with 6 servers and business process servers, an online backup server with 40TB capacity, along with supporting servers that provide print services, LDAP, DHCP, DNS, tapes backup and software distribution to the department. The data center is internally connected via Gigabit Ethernet and accesses the Internet and Internet2 via the campus fiber network. The University provides 802.11b coverage for most of the campus.

Columbia University Information Technology maintains Columbia's commodity Internet connections that include 1 Gb/s via Level 3, 5 Gb/s via Cogent and 100 Mb/s via RCN (for RCN residential cable modem customers). Columbia's research and education network connections include a NYSERNet NYC and statewide regional optical network ("dark fiber"), a 1 Gb/s connection to NYSERNet's IP network and from there to Internet2 and National Lambda Rail (NLR) packet net and 1 Gb/s connection to the US Large Hadron Collider Network (USLHCNET) providing connectivity to CERN.

The research infrastructure includes many Sun, Dell and HP servers and hundreds of workstations from Apple, Dell, HP and other PC manufacturers. The labs for research in image processing, vision, graphics, and robotics contain such specialized equipment as a Data Cube image processor, an Adept1 robot, and Aspex PIPE (an eight-stage parallel, pipelined, low-level image processor), Unimate Puma 500 and IBM robotic arms, a Utah-MIT dextrous hand, high performance 3-D graphics workstations, 3-D position/orientation trackers, see-through headworn displays, and a wall-sized stereo display.

CRF facilities are partially supported by the Secure Cyber Operations and Parallelization Studies (SCOPS) instrumentation grant from AFOSR (Contract #: FA 99500910389).

Technical Support

The research facilities are staffed by professional systems engineers who are responsible for operating system and network support, miscellaneous hardware and software maintenance, and troubleshooting. These staff members allow individual researchers to avoid spending time on hardware and software problems.