## Facilities, Equipment and Other Resources

**Frost Institute for Data Science and Computing (IDSC)**

The University of Miami Frost Institute for Data Science and Computing (IDSC) is focused on tapping the extraordinary potential of data science and computing for interdisciplinary and convergent research and innovation. IDSC takes a collaborative approach to bring together talented minds in multiple disciplines and leverages the Institute’s state-of-the-art technology including an AI-ready supercomputer and massively parallel computing capabilities. . IDSC enables discovery through data-intensive research in diverse fields across science, engineering, liberal arts, and medicine. Ongoing research and engagement include medicine to earth sciences, urban planning, digital humanities, and business. IDSC also strives to enhance data science understanding among its students and the public, helping individuals of all ages, including underrepresented minorities, develop the STEM skills needed for professional success in the 21st century.

As a member of the University of Miami’s Frost Institutes of Science and Engineering, IDSC supports basic and applied research initiatives throughout the enterprise and with other leading institutions around the world. IDSC is committed to developing strong partnerships with industry, government, and nonprofit institutions in south Florida, while building bridges to the Caribbean, Latin America, and beyond.

**IDSC RESEARCH PROGRAMS**

*Artificial Intelligence + Machine Learning*

AI encompasses machine learning (ML), natural language understanding (spoken/written), computer vision, data mining, human-computer interfaces, data visualization, and deep learning. The application of AI and ML is ushering in crucial advances in better understanding hidden theories of science, medicine, business, and even the arts, humanities, and social sciences, assisting humans in decision making and in solving complex problems expeditiously. The AI + ML research group provides expertise and capabilities to further explore high-dimensional data.

*Data Ethics + Society*

In conjunction with UM’s Ethics Programs and the Institute for Bioethics and Health Policy, IDSC is identifying, addressing, and resolving research challenges from the appropriate uses and users of intelligent machines to privacy challenges raised by data collection, analysis, surveillance, and secondary use.

*Digital Health + Life Sciences Informatics*

With access to massive amounts of structured and unstructured patient data across a wide range of data sources, IDSC data-driven exploration in digital drug discovery, population health informatics, precision medicine, and social and behavioral data science can aid in diagnosis, developing new therapeutics, matching treatment with best outcomes, and predicting patient risk levels for disease. UM is well positioned to harness the power of the data of millions of patients, combining impressive academic programs with one of the most extensive health care systems in the State.

*Earth Systems Science*

Machine learning (ML) and big data analytics are highly effective tools for developing models and making predictions in Earth Systems Science where data are sparse, and uncertainty high. The pairing of ML with Bayesian statistics is rendering multiple prediction tools that provide more reliable forecasts and a detailed understanding of inherent uncertainties in extreme weather events.

*Human Centered Design + Computing*

IDSC Human Centered Design and Computing, encompassing Data Visualization and Creative Technologies, fulfills a key educational role in raising awareness about data science and its applications. The use of multimodal media—from static infographics to interactive and XR technologies—helps students and scientists illuminate their data and communicate their findings. Its XR Studio contains resources critical to extended reality development.

*Smart Cities + Smart Environments*

Smart Cities research tackles data-collecting technology and data analytics, and innovation and utilization in both hardware and software applications. This combination leverages the U’s computational resources and expertise in physical computing and data analytics to design the next generation of smart cities and environments.

*Advanced Computing*

Addressing the ever-expanding needs of data-driven research, the Triton supercomputer, a GPU-accelerated system, represents a quantum leap in the University of Miami’s computing infrastructure. Built using IBM Power Systems AC9222 servers, Triton was designed to maximize data movement between the IBM POWER9 CPU and attached GPU accelerators, and to accommodate traditional high-performance computing, interactive data science, big data, AI, and ML workloads.

*Systems + Data Engineering*

The Systems and Data Engineering team are professional software engineers who actively seek collaborative partners for new and innovative software application and systems development projects. The team has developed applications and software systems to support work in clinical research, drug discovery, genomics, mapping, urban planning, observational biology/ecology, and the digital humanities.

*Office of Industrial Partnerships (OIP)*

IDSC’s collaborative approach brings together talented minds at the interface of disciplines. From creative start-ups to ideas for industry-shaping innovations, IDSC’s Office of Industrial Partnerships (OIP) leads the development of innovation partnerships with local and international stakeholders from business, industry, and the broader community. Integral to these efforts are building collaborations that leverage the University’s internationally recognized faculty and research staff, along with its world-class computing infrastructure. The OIP works closely with IDSC Industry Advisory Board who is committed to making South Florida a hub for the integration of data science into the decision-making processes for all industries; and also, the Office of Technology Transfer to protect and commercialize innovations from these partnerships.

*Administration and Business Office*

The IDSC Business Office provides the administrative framework to ensure the smooth and efficient operation of the Institute. The office directs all fiscal, facilities, operational, marketing and communication initiatives, and strategic recruiting activities. The business office also provides project management, pre and post award administration, and IRB support. The office works closely with IDSC’s Office of Industrial Partnerships and liaises with other key administrators across all three university campuses.

**FACILITIES**

**Offices**

University of Miami Frost Institute for Data Science and Computing (IDSC) faculty and staff hold offices on all three campuses. Each location is equipped with workstations and essential software applications.

IDSC has three dedicated conference rooms and communication technology to interact with advisors (telephone, web-, and videoconferencing), plus a [Visualization Lab](https://idsc.miami.edu/visualization-lab/) with 2D and 3D display walls.

**NAP of the Americas**

IDSC systems are co-located at the [Century Link Data Center](https://www.centurylink.com/business/colocation/data-centers/miami.html) hosted by the Equinix NAP of the Americas (NOTA or "the NAP"). The NAP in Miami currently features a 750,000-square-foot, purpose-built data center Tier IV facility with N+2 14-Megawatt power and cooling infrastructure. The equipment floors start at 32 feet above sea level, and the roof slope was designed to aid in the drainage of floodwater in excess of 100-year-storm intensity, assisted by: 18 rooftop drains, architecture designed to withstand a Category 5 hurricane with approximately 19 million pounds of concrete roof ballast, and 7-inch-thick steel reinforced concrete exterior panels. Plus, the building is outside FEMA’s 500-year designated flood zone. The NAP uses a dry pipe fire-suppression system to minimize the risk of damage from leaks.

The NAP features a centrally located Command Center manned by 24x7 security and security sensors. In order to connect the University with the NAP, UM has invested in a Dense Wavelength Division Multiplexing (DWDM) optical ring for all campuses. IDSC systems occupy a discrete, secure wavelength on the ring, which provides a distinct 10 Gigabit (GB) advanced computing network to all campuses and facilities.

Given the University’s past experiences (including several hurricanes and other natural disasters), we anticipate no service interruptions due to facilities issues. The NAP was designed and constructed for resilient operations. The University has gone through several hurricanes, power outages, and other severe weather crises without any loss of power or connectivity to the NAP. The NAP maintains its own generators with a flywheel power crossover system. This ensures that power is not interrupted when the switch is made to auxiliary power. The NAP maintains a two-week fuel supply (at 100% utilization) and is on the primary list for fuel replacement due to its importance as a data-serving facility.

In addition to hosting the University of Miami’s computing infrastructure, the NAP is home to the US SouthCOM, Amazon, eBay, and several telecommunications companies’ assets. The NAP at Miami hosts 97% of the network traffic between the US and Central/South America and is the local access point for Florida Lambda Rail (FLR)—which is gated to Internet 2 (I2) to provide full support to the I2 Innovation Platform. The NAP also provides TLD information to the DNS infrastructure and is the local peering point for all networks in the area.

The [University of Miami](http://www.welcome.miami.edu/) has made the NAP its primary Data Center occupying a very significant footprint. Currently all IDSC resources, clusters, storage, and backup systems run from this facility, and serve all major campuses.

**EQUIPMENT**

The University is powered by the vast resources of IDSC, which has an established, broad user base, and a state-of-the-art research computing infrastructure

#### Advanced Computing

Through IDSC, the University of Miami maintains one of the largest centralized academic cyber infrastructures in the country with numerous assets. Since 2007, the core has grown from zero advanced computing cyberinfrastructure to a regional high-performance computing environment that currently supports more than 50+ user groups spanning across centers and institutes, 240 Teraflops (TFLOPS) of computational power, and more than 3 Petabytes (PB) of disk storage.

#### Triton Supercomputer

The University’s latest acquisition, “Triton,” was rated one of the **Top 5 Academic Institution Supercomputers in U.S. for 2019 and** is UM’s first GPU-accelerated advanced computing (AC) system, representing a completely new approach to computational and data science for the University. Built using **IBM Power Systems AC922 servers**, this system was designed to maximize data movement between the **IBM POWER9 CPU** and attached accelerators like GPUs.

#### Triton Specs

* IBM Power9/Nvidia Volta – 6 Racks
* IBM Declustered Storage – 2 Racks
* 96 IBM Power 9 servers
* 30 Terabytes (TB) RAM (256/node)
* 1.2 Petaflop (PFLOPS) Double Precision
* 240TFLOPS Deep Learning
* 64-bit Scalar
* 100 GB/sec Storage
* 150 TB Shared Flash Storage
* 400 TB Shared Home
* 2 @ 1.99 TB SSD Local Storage

#### Pegasus Supercomputer

When "Pegasus" (the University’s first supercomputer) was launched, it ranked at number 389 on the [November 2012 Top 500 Supercomputer Sites list](http://www.top500.org/list/2012/11/). Pegasus is a 350-node Lenovo cluster with each node having 2 Intel Sandy Bridge E5-2670 (2.6 GHz) 8C—with 32 GB 1600 MHz RAM (2 GB/core) for a total of over 160 TFlops. Connected with an FDR InfiniBand fabric, Pegasus was purpose-built for the style of data processing performed by biomedical research and analytics. In contrast with traditional supercomputers where data flows along the slowest communication network possible (Ethernet), Pegasus was built on the principle that data needs to be on the fastest fabric possible. By utilizing the low latency high bandwidth IB fabric for data, Pegasus was able to access all three tiers (SSD, 15K RPM SAS, 7.2K NL-SAS) at unprecedented speeds.

Unlike traditional advanced computing storage, the 150 TB /scratch filesystem is optimized for small random reads and writes; and can support over 125,000 sustained IOPs/second and 20 Gb/sec throughput at 4Kb file size. Composed of over 500 15K RPM SAS disks, /scratch is ideal for the extremely demanding IO requirements of biomedical workloads.

For instances where even /scratch is not fast enough, Pegasus has access to over 8TB of burst buffer space clocked at over 1,000,000 IOPs. This buffer space provides biomedical researchers a good place for large file manipulation and transformation.

Along with the 350 nodes in the general processing queue, all researchers also have access to the 20 large memory nodes in the bigmem queue. With access to the entire suite of software available on Pegasus, the bigmem queue provides large memory access (256 GB) to researchers where parallelization is not an option. With 20 cores each, the bigmem servers provide an SMP-like environment well suited to biomedical research.

As many modern analysis tools require interaction, Pegasus has a unique feature of allowing SSH and graphical (GUI) access to programs using LSF. Tools ranging from MATLAB to Knime and SAS to R are available to researchers in the interactive queue with full-speed access to /scratch and the W.A.D.E. Storage Cloud.

#### **Pegasus Specs**

Five Racks of iDataPlex in iDataPlex Racks

* One Standard Enterprise Rack for Networking and Management
* iDataPlex dx360 M4:
	+ Qty (2) Intel Sandy Bridge E5-2670 (2.6 GHz)- 32 GB 1600 MHz RAM (2GB/core)
	+ Stateless/Diskless
	+ Mellanox Connect X3 Single-Port FDR
* Mellanox FDR MSX6036
* DNA SFA 12k:
	+ Qty (12) 3TB 7.2K RPM SATA (RAID 6 in 8+2)
	+ Qty (360) 600GB 15K SAS (RAID 6 in 8+2)
	+ Qty (10)e 400GB MLC SSD (RAID 1 Pairs)
* xCAT 2.7.x
* Platform LSF
* RHL 6.2 for Login/Management Nodes

Pegasus’ CPU Workhorse—The IBM iDataPlex dx360M4

**Compute Nodes** 350 dx360 M4 Compute Nodes
**Processor** Two 8-core Intel Sandy Bridge 2.6 GHz scalar, 2.33 GHz\* AVX
**Memory** 32 GiB (2 GiB/core) using eight x 4GB 1600MHz DDR3 DIMMs
**Clustering Network** One FDR InfiniBand HCA
**Management Network** GB Ethernet NIC connected to the cluster management VLANs. IMM access shared through the eth0 port

#### W.A.D.E. Storage Cloud (Worldwide Advanced Data Environment)

At the heart of the IDSC Advanced Computing data services is the W.A.D.E. Storage Cloud, which currently provides over 7 PB of active data to the University of Miami research community ranging from small spreadsheets in sports medicine research, to multi-terabyte, high-resolution image files and NGS datasets. W.A.D.E. is composed of four DDN storage clusters running the GPFS filesystem. The combination of IBM’s industrial strength filesystem and DDN’s high performance hardware gives researchers at UM the flexibility to process data on Pegasus and share that data with anyone, anywhere.

By utilizing several file-service gateways, researchers can share large data sets securely between Mac, Windows, and Linux operating systems. Data can also be presented outside of the University in several high-performance ways. In addition to the common protocols of SCP and SFTP, we also provide high-speed parallel access through bbcp and Aspera. You can even share your data using standard web access (httpd) through our integrated web and cloud client service.

All access to W.A.D.E. is provided through UM’s 10 GB/sec Research Network internally and the UM Science DMZ externally. All Internet traffic flows through either the Science DMZs 10 GB/Sec I2 link through Florida Lambda Rail or through the Research Network’s 1 GB/sec commercial internet connect.

#### Vault Secure Storage Service

The Vault Secure Storage Service is designed to address the ongoing challenge of storing Limited Research Datasets. Built on enterprise-quality hardware with 24x7 support, Vault provides CTSI-approved researchers access to over 150TB of usable redundant (300 TB raw) storage. All data is encrypted according to U.S. Federal Information Processing Standards (FIPS). At rest, data is encrypted using AES encryption with 128-bit keys. In motion, all transfers are encrypted using FIPS 140-2 compliant AES with 256-bit keys. All data is encrypted and decrypted on access automatically.

Access to the Vault storage service is controlled through several methods including the latest in multifactor authentication. All users are required to use YubiKey ™ 4 hardware USB keys in order to log on to the vault secure storage service. Vault also requires IP whitelisting for access through either the on-campus research network or campus-based VPN services.

#### Visualization Laboratory

The Visualization Laboratory (Viz Lab) is a tool for all University of Miami students and faculty to present graphical and performance intensive 2D and 3D simulations. With a direct connection to all University Advanced Computing resources, the Viz Lab is the perfect tool for high-performance parallel visualization, data exploration, and other advanced 2D and 3D simulations.

The Viz Lab is built around a Cyviz 5x2 20 Megapixel Native display wall and Mechdyne 2x2 passive 3D display wall. On these impressive high-resolution displays, users can present their work at a paramount level while analyzing details at a granular level.

The Viz Lab sits directly on the Research Network, providing 10GB/sec network access to the Storage Cloud, Triton and Pegasus Supercomputers, and all other Advanced Computing resources. It was built with the focus to interpret real-world scenarios such as computational modelling, simulation, analysis, visualization of natural and synthetic phenomena for dynamic engineering, biomedical, epidemiological, and geophysical applications.

The 2D display is composed of ten 55-inch, thin bezel, LCD Planar panels spanning 22 feet for an ultra-wide-angle 21-megapixel display that supports a resolution of up to 9600 x 2160.

The 3D display wall supports stereoscopic 3D, for users looking to captivate audiences with something a little more eye-popping or simply looking to add depth to their work. It is composed of four 46-inch ultra-thin LCD Planar panels and supports resolutions up to 5120 x 2880.

#### Secure Processing Service (SPS)

Our most secure data processing offering is SPS, designed for secure access to extremely sensitive data sets including PHI. In addition to the security protocols used in the Vault data services, SPS requires additional administrative action for the certified placement and/or destruction of data. IDSC Advanced Computing staff (all CITI trained and IRB approved) act as data managers for several federal agencies including NSF, NIH, DoL, DoD, and VA projects. Once our staff has loaded and secured your data, you can remotely access one of the SPS servers (either Windows or Linux) which has access to the most common data analytic tools including R, SAS, MATLAB, and Python. Additional tools are available on request. 50 TB of highly secure redundant storage (100 TB raw).

#### NYX Cloud

The NYX Cloud hosting system allows launch and configuration of your own Virtual Machine servers. It is a private UM Cloud system powered by the OpenStack cloud software, which offers the IaaS (Infrastructure-as-a-Service) resource management. It is available to registered users, and resource allocations are project-based.

NYX Cloud Virtual machine (VM) instances are grouped into projects, which reside on dedicated private virtual networks (subnets). Through the dashboard, users can start and customize their own VMs. NYX VMs can be single- or multiple-CPU servers and can be shut down or restarted as needed. Several bootable images are available, including configurations such as LAMP and MEAN in CentOS 6 and CentOS 7. Snapshots of projects can be taken for replication and backup. Floating IP addresses are available for SSH connections to Nyx instances from outside the project network.

Advanced storage features include block and object storage. VM instances can be started with dedicated block storage or attached to existing block storage. Object (distributed) storage, which allows for data access via HTTP, is also available.

**OTHER RESOURCES**

### **EXPERTISE**

The IDSC Advanced Computing team has in-depth experience in various scientific research areas with extensive experience in parallel and distributed high performance computing codes written in Fortran, C, Java, Perl, Python and R. The team is active in contributing to Open-Source software efforts including R, Python, the Linux Kernel, Torque, Maui, XFS and GFS. The team also specializes in scheduling software (LSF) to optimize the efficiency of the advanced computing systems and adapt codes to the IDSC environment. The Advanced Computing team has expertise in message-passing interface (MPI) and OpenMP depending on the programming paradigm. IDSC has contributed several parallelization efforts back to the community in projects such as R, WRF, and HYCOM.

The Advanced Computing environment currently supports more than 300 applications and optimized libraries. Experts in implementing and designing solutions in the three different variants of Unix, the Advanced Computing team also maintains industry research partnerships with IBM, Schrodinger, Open Eye, and DDN.

### **SOFTWARE**

The Advanced Computing team continually updates applications, compilers, system libraries, etc. To facilitate this task and to provide a uniform mechanism for accessing different revisions of software, the team uses the modules utility. At login, modules commands set up a basic environment for the default compilers, tools, and libraries such as: the $PATH, $MANPATH, and $LD\_LIBRARY\_PATH environment variables.

Users are free to install software in their home directories. Requests for new software are reviewed quarterly. Global software packages are considered when a minimum of 20 users require them.

### **Artificial Intelligence and Machine Learning**

The AI + ML research group provides expertise and capabilities to further explore high-dimensional data. The following are examples of the expertise areas covered by our team:

* Classification, which appears essentially in every subject area that involves collection of data of different types, such as disease diagnosis based on clinical and laboratory data. Methods include regression (linear and logistic), artificial neural nets (ANN), k-nearest neighborhood (KNN), support vector machines (SVM), Bayesian networks, decision trees and others.
* Clustering, which is used to partition the input data points into mutually similar groupings, such that data points from different groups are not similar. Methods include KMeans, hierarchical clustering, and self-organizing map (SOM), and are often accompanied by space decomposition methods to offer low dimensional representations of high dimensional data space. Methods of space decomposition include principal component analysis (PCA), independent component analysis (IDA), multidimensional scaling (MDA), Isomap, and manifold learning. Advanced topics in clustering include multifold clustering, graphical models, and semi-supervised clustering.
* Association data mining, which finds frequent combinations of attributes in databases of categorical attributes. The frequent combinations can be then used to develop prediction of categorical values.
* Analysis of sequential data involves mostly biological sequence and includes such diverse topics as extraction of common patterns in genomic sequences for motif discovery, sequence comparison for haplotype analysis, alignment of sequences, and phylogeny reconstruction.
* Text mining, particularly in terms of extracting information from published papers, thus transforming documents to vectors of relatively low dimension to enable the use of data mining methods mentioned above.

### **Drug Discovery**

IDSC has a sophisticated cheminformatics and compute infrastructure with a significant level of support from the institution. IDSC facilitates scientific interactions and enables efficient research using informatics and computational approaches. A variety of departments and centers at the University use high content and high throughput screening approaches—[The Miami Project to Cure Paralysis](http://www.themiamiproject.org/), [Diabetes Research Institute](http://www.diabetesresearch.org/diabetes-research-institute), [Sylvester Comprehensive Cancer Care Center](http://sylvester.org/?gclid=CJSHwOmhxM4CFUFahgodr0sOZw), [Bascom Palmer Eye Institute](http://bascompalmer.org/), the [Department of Surgery](http://surgery.med.miami.edu/), and [John P. Hussman Institute for Human Genomics](http://hihg.med.miami.edu/).

Cheminformatics and computational chemistry tools—running on Advanced Computing's Linux cluster and high-performance-application server:

* SciTegic Pipeline Pilot—visual work-flow-based programming environment (data pipelining); broad cheminformatics, reporting / visualization, modeling capabilities; integration of applications, databases, algorithms, data.
* Leadscope Enterprise*—*integrated cheminformatics data mining and visualization environment; unique chemical perception (~27K custom keys; user extensions); various algorithms, HTS analysis, SAR / R-group analysis, data modeling.
* ChemAxon Tools and Applications—cheminformatics software applications and tools; wide variety of cheminformatics functionality.
* Spotfire*—*highly interactive visualization and data analysis environment, various statistical algorithms with chemical structure visualization, HTS and SAR analysis.
* Open Eye ROCS, FRED, OMEGA, EON, etc. implemented on Linux cluster - suite of powerful applications and tool kits for high-throughput 3D manipulation of chemical structures, modeling of shape, electrostatics, protein-ligand interactions and various other aspects of structure- and ligand-based design; also includes powerful cheminformatics 2D structure tools.
* Schrodinger Glide, Prime, Macromodel, and various other tools implemented on Linux Cluster—powerful state of the art docking, protein modeling and structure prediction tools and visualization.
* Desmondimplemented on Linux Cluster—powerful state of the art explicit solvent molecular dynamics.
* TIP workgroup*—*powerful environment for global analysis of protein structures, binding sites, binding interactions; implemented automated homology modeling, binding site prediction, structure and site comparison for amplification of known protein structure space.

### **Visualization, Data Communication, and Information Design**

The Information Design program conducts both theoretical and applied research in the general areas of Machine Vision and Learning, and specifically in:

* Computer Vision and Image Processing
* Machine Learning
* Biomedical Image Analysis
* Computational Biology and Neuroscience

The goal is to provide expertise in this area to develop novel fully automated methods that can provide robustness, accuracy and computational efficiency. The program works towards finding better solutions to existing open problems in the above areas, as well as exploring different scientific fields where our research can provide useful interpretation, quantification and modeling.

**XR Studio**

The XR Studio contains resources critical to extended reality development. Specifically, the studio has three (3) Alienware Aurora workstations (Intel (R) Core i9-9900K CPU @ 3.60GHz; NVIDIA GeoForce RTX 2080 Ti; 64 GB RAM). Studio workstations are equipped with the Unity game engine for creating XR experiences and Blender for creating any necessary 3D models. The studio uses HP Reverb G2 Omnicept Edition virtual reality headsets with dual LCD 2.89’’ diagonal with Pulse Backlight technology, 2160 x 2160 pixels per eye resolution with a 90 Hz refresh rate and ~114 degrees, Fresnel-Aspheri field of view. The headsets can track biometrics (eye tracking, pupillometry, heart rate, heart rate variability, and cognitive load) from users wearing the device in real-time.