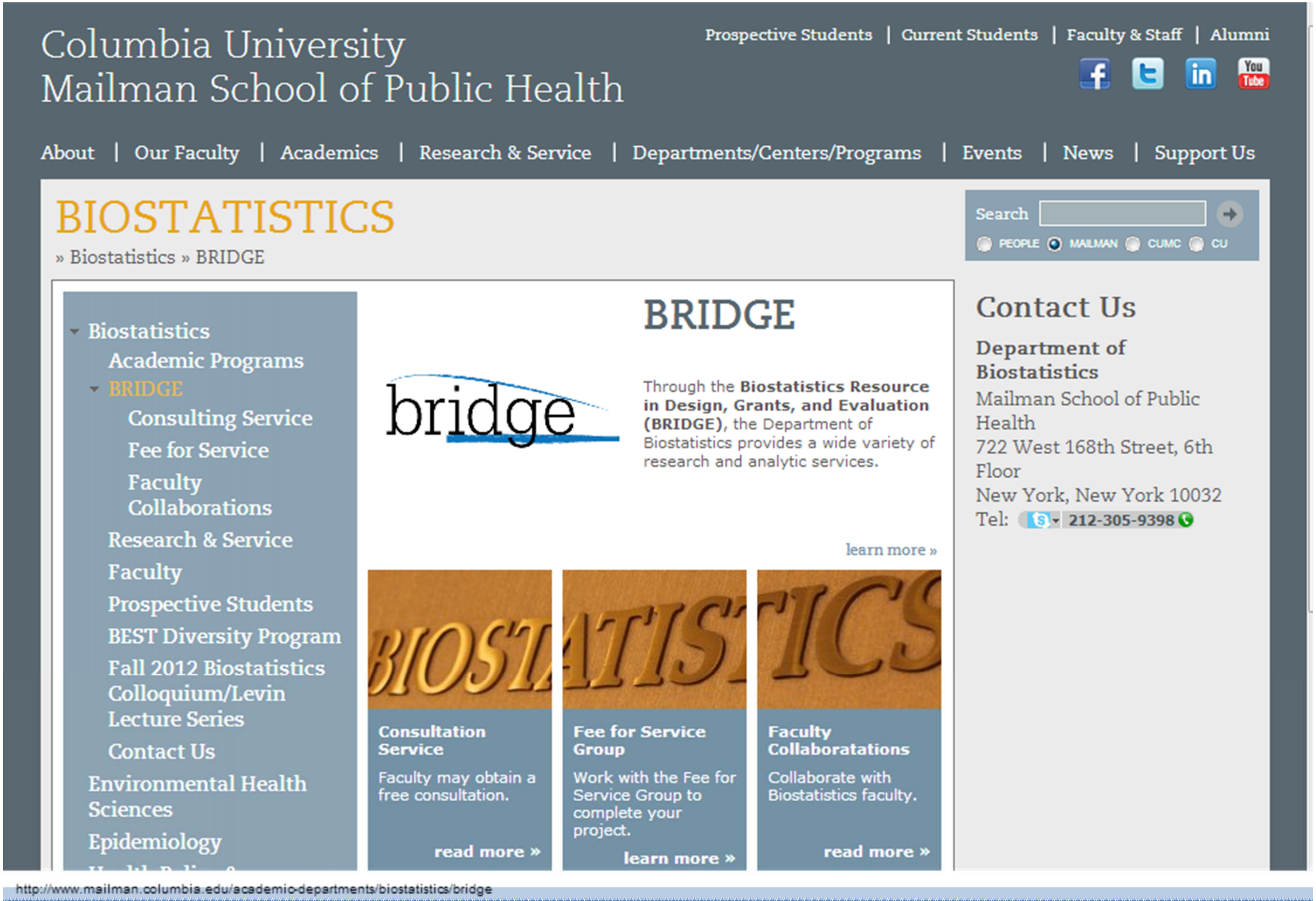
High Performance Computing

June 11th, 2013

What do you need ?

1. Database Management?
2. Statistical Analysis?
3. HPC?



BIOSTATISTICS

» Biostatistics » BRIDGE

BRIDGE

Through the **Biostatistics Resource in Design, Grants, and Evaluation (BRIDGE)**, the Department of Biostatistics provides a wide variety of research and analytic services.

learn more »

Consultation Service

Faculty may obtain a free consultation.

read more »

Fee for Service Group

Work with the Fee for Service Group to complete your project.

learn more »

Faculty Collaboratations

Collaborate with Biostatistics faculty.

read more »

Contact Us

Department of Biostatistics
Mailman School of Public Health
722 West 168th Street, 6th Floor
New York, New York 10032
Tel: 212-305-9398

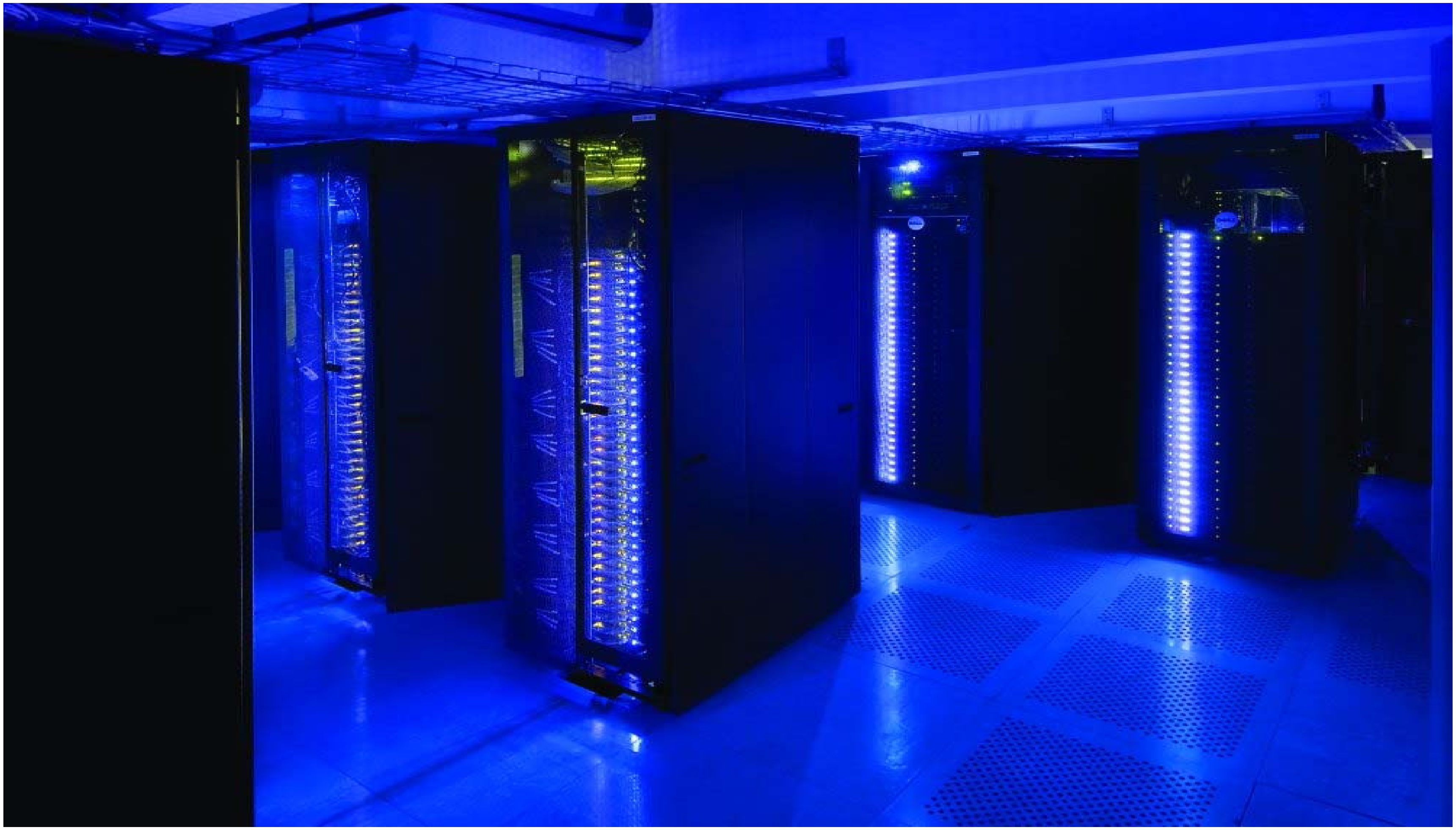
http://www.mailman.columbia.edu/academic-departments/biostatistics/bridge

High Performance Computing

- What is it ?
- **Physically**: it is not simply a bigger desk top computer; we speak in Terabytes and Teraflops.
- **Conceptually**: HPC is an essential research tool for every operationalization of Big Data (i.e. Systems Science, GIS mapping, Simulation Sciences, NextGen Sequencing Analysis, any –Omics, any Mash-Up, functional analysis, analysis of images).
- HPC also allows us to venture into, and take our research into, competitive new space, and the cutting edge unknown.

Floppy AND a 10Mb hard drive?

High Performance Computing (HPC)

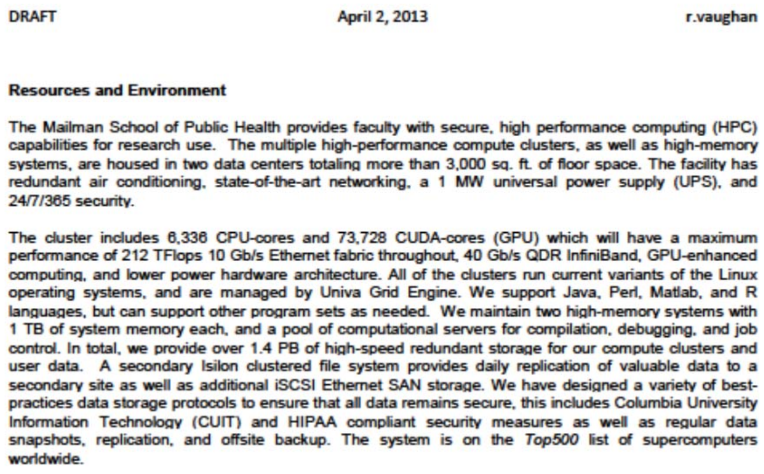


High Performance Computing

- To keep us competitive, the **Dean is now able to provide HPC for faculty (and supervised students)**, in collaboration with the system at C2B2 (now ranked one of the top 500 supercomputers in world)!
- Over 6,000 compute cores (CPU), over 70,000 (GPU's), running at >200 Tflops, with nearly 2PB of storage.
- Translation ? Really Big and Fast and Powerful.

HPC Roll Out

- We have hired a MSPH- C2B2 Research Computing Liaison (Ms. Rebecca Yohannes) to assist with accounts, troubleshooting, FAQ's.
- We will organize HPC working groups, hold trainings & workshops, and HPC presentations.
- Will have regular presentations on "tips" through R^2. An NIH style "Resources and Environment" description on HPC is now available through R^2.
- Year 1 is trial year. Chargeback plans / writing costs into grants forthcoming.

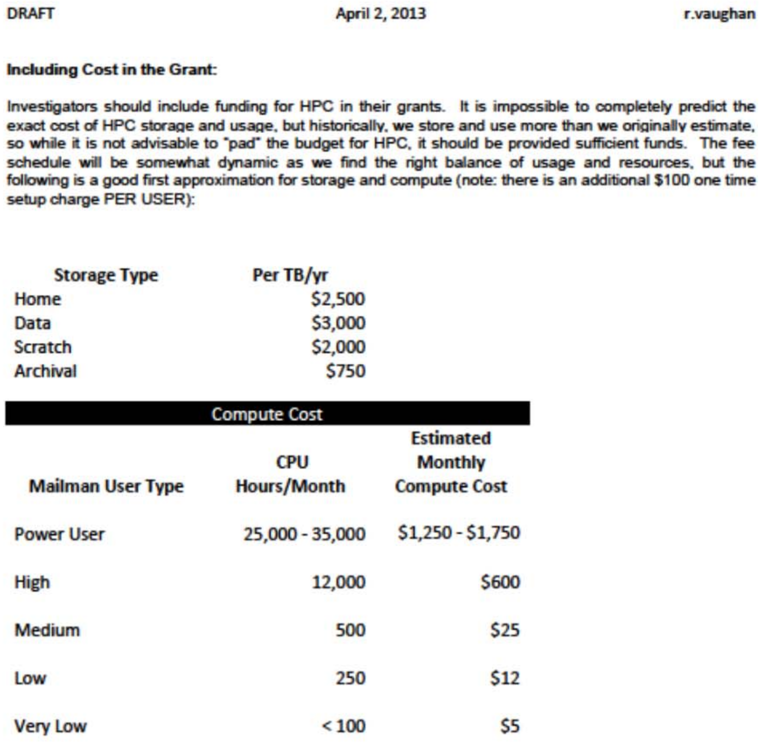


DRAFT April 2, 2013 r.vaughan

Resources and Environment

The Mailman School of Public Health provides faculty with secure, high performance computing (HPC) capabilities for research use. The multiple high-performance compute clusters, as well as high-memory systems, are housed in two data centers totaling more than 3,000 sq. ft. of floor space. The facility has redundant air conditioning, state-of-the-art networking, a 1 MW universal power supply (UPS), and 24/7/365 security.

The cluster includes 6,336 CPU-cores and 73,728 CUDA-cores (GPU) which will have a maximum performance of 212 TFlops 10 Gb/s Ethernet fabric throughout, 40 Gb/s QDR InfiniBand, GPU-enhanced computing, and lower power hardware architecture. All of the clusters run current variants of the Linux operating systems, and are managed by Univa Grid Engine. We support Java, Perl, Matlab, and R languages, but can support other program sets as needed. We maintain two high-memory systems with 1 TB of system memory each, and a pool of computational servers for compilation, debugging, and job control. In total, we provide over 1.4 PB of high-speed redundant storage for our compute clusters and user data. A secondary Isilon clustered file system provides daily replication of valuable data to a secondary site as well as additional iSCSI Ethernet SAN storage. We have designed a variety of best-practices data storage protocols to ensure that all data remains secure, this includes Columbia University Information Technology (CUIT) and HIPAA compliant security measures as well as regular data snapshots, replication, and offsite backup. The system is on the *Top500* list of supercomputers worldwide.



DRAFT April 2, 2013 r.vaughan

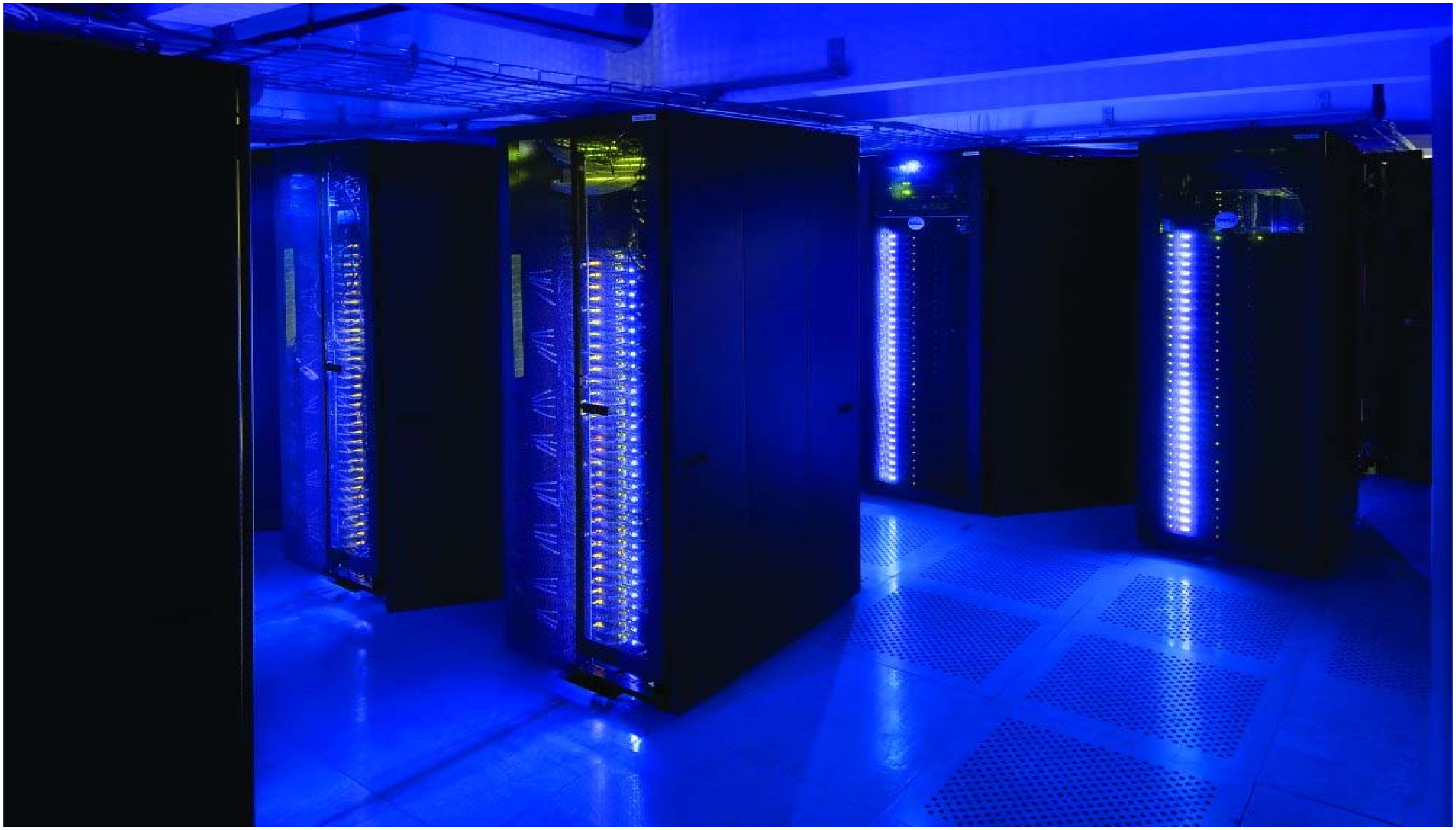
Including Cost in the Grant:

Investigators should include funding for HPC in their grants. It is impossible to completely predict the exact cost of HPC storage and usage, but historically, we store and use more than we originally estimate, so while it is not advisable to "pad" the budget for HPC, it should be provided sufficient funds. The fee schedule will be somewhat dynamic as we find the right balance of usage and resources, but the following is a good first approximation for storage and compute (note: there is an additional $100 one time setup charge PER USER):

Storage Type	Per TB/yr
Home	$2,500
Data	$3,000
Scratch	$2,000
Archival	$750

Compute Cost		
Mailman User Type	CPU Hours/Month	Estimated Monthly Compute Cost
Power User	25,000 - 35,000	$1,250 - $1,750
High	12,000	$600
Medium	500	$25
Low	250	$12
Very Low	< 100	$5

High Performance Computing (HPC)



How can one benefit from using the High Performance Computing Cluster?

- Run multiple iterations of a model all at once
  - Parallel computing
  - Scalability

- Availability of a wide variety of research software packages

- Jobs that take days to run can be submitted to the HPC freeing up personal computer

What does one need to get started?

- Access to the HPC (userid/password)
  - Need to attend training
- SSH (Secure Shell) to the HPC
- Basic knowledge of Unix and shell scripting

Available resources-computing/software/storage

•The latest GNU and Intel compilers for C and Fortran, Perl interpreters, Java SDKs

•Popular bioinformatics and statistics software and environments like Matlab, BLAST, EMBOSS, HMMER, MUMmer, clustalW, PAML, PHYLIP, BioConductor, Phred and Phrap, GeneHunter, Fastlink, Merlin, PDT, TRANSMIT, Pseudomarker, Analyze, Autosacan, GOLD, plus many other utilities and programs

•Petabyte disk storage with regular back up system

Support and Training

- Determine each project's requirements,

- Design a custom solution that encompasses:
  - software environment,
  - automation of tasks and
  - operational support.

How to get access?

- Send a request to

  ry111@cumc.columbia.edu with:

  - Full name
  - Columbia UNI
  - Name of Department
  - Project name and
  - Project Description