MARCC: On Data Subject to Restrictions

Jaime E. Combariza
Associate Research Professor
Department of Chemistry
Director MARCC
Johns Hopkins University

2016 IDIES Annual Symposium

MARCC

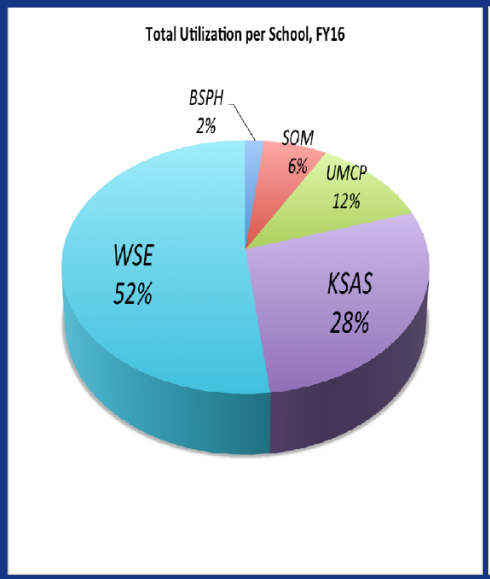
Configuration 2015

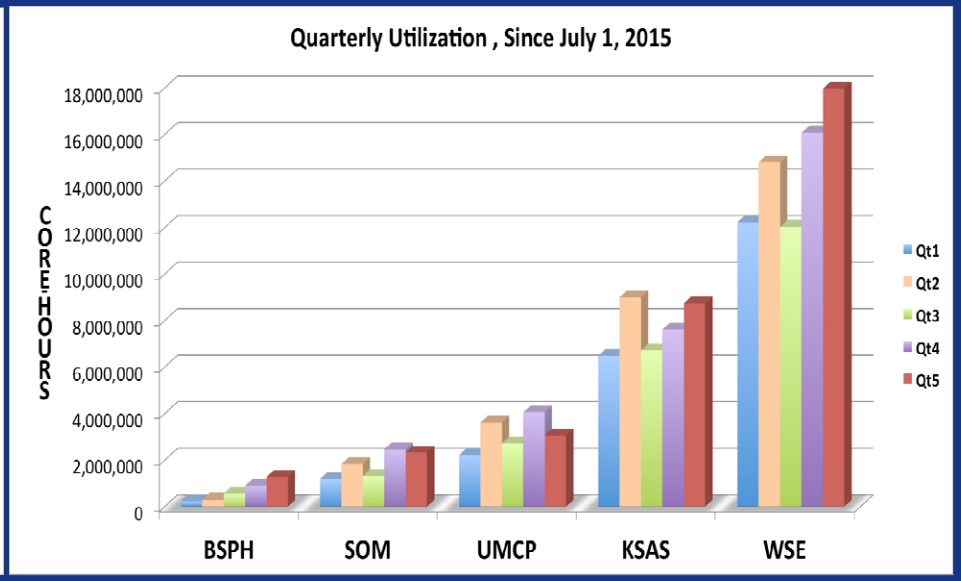
JOHNS HOPKINS UNIVERSITY

Count	Description
648	Compute nodes, 128GB RAM, 24 cores, 2.5GHz Haswell processors
50	Large memory nodes, 1024 GB RAM, 48 cores
48	GPU nodes, 2 Nvidia K80 GPUs/node, 24 CPU cores
2 PBytes	High Performance File System (Lustre)
14 Pbytes	ZFS File System

Condo, 28 nodes (MEDE + 2 PIs) :: 19,776 cores 932.5 TFLOPs

Utilization





Metrics

Number of Research Groups	> 200
Number of Users	> 950
Number of CPU-cycles	106,299,604
Effective Utilization	~85%
Number of Scientific Applications	> 300
Number of Publications	> 70
Number of Classes	10
Training	>600 users

GPU Computing

NWChem	TeraChem	Quantumespresso	Gamess
VASP	LAMMPS	Abaqus	Terachem
Amber	NAMD	Gromacs	Ab Init
Caffe	Torch7	Theano	Charmm
Lsdyna	Bluebottle	S3D	WRF
NVBowtie	Soap3	GPUBlast	Matlab

MARCC

Configuration 2016

JOHNS HOPKINS UNIVERSITY

Count	Description
676	Compute nodes, 128GB RAM, 24 cores, 2.5GHz Haswell processors
50	Large memory nodes, 1024 GB RAM, 48 cores
48	GPU nodes, 2 Nvidia K80 GPUs/node, 24 CPU cores
+48	Compute nodes, 128GB RAM, 28 cores, 2.6 Broadwell Processors
+24	GPU nodes, 2 Nvidia K80 GPUs, 28 cores.

Bluecrab 2016 :: 21,792 cores; 1.1 PFLOPs

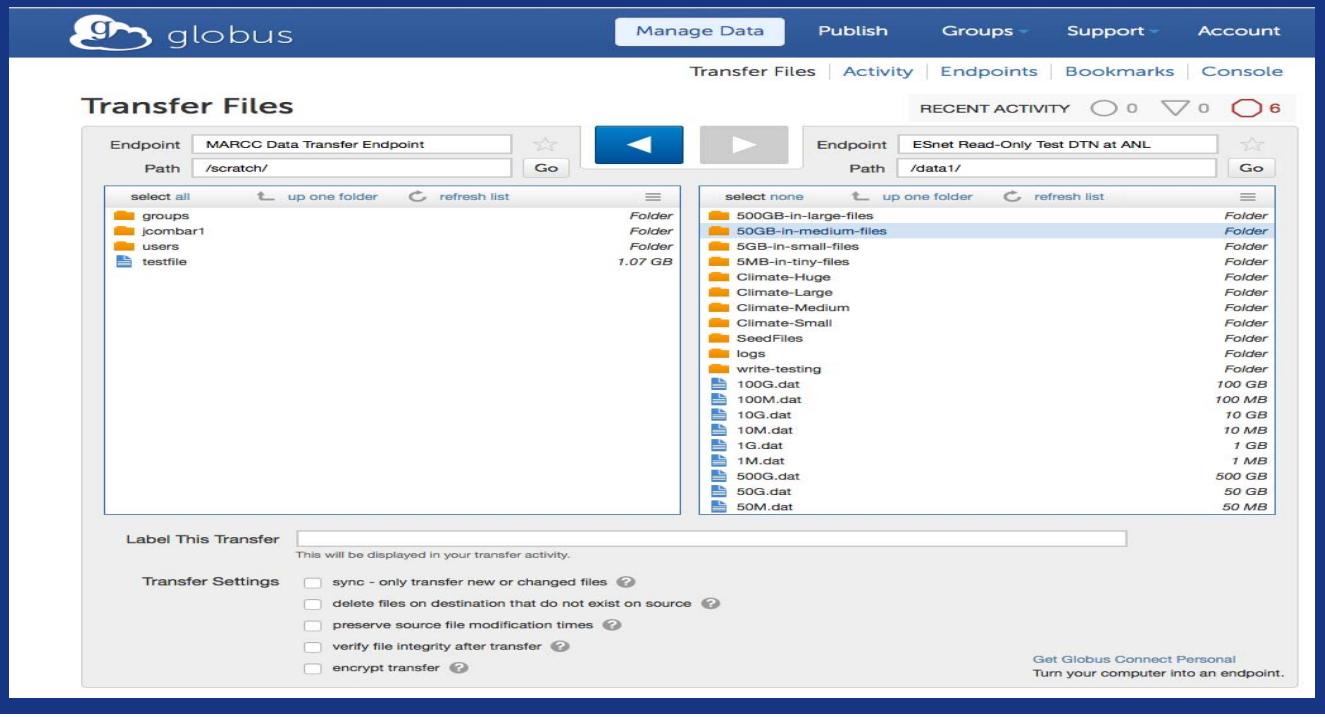
MARCC

What is new

JOHNS HOPKINS UNIVERSITY

- Two Factor Authentication (TFA)
  - Gloogle authenticator
- Security (Bro-IDS, IPTables)
  - Log/network traffic analysis
- Enhanced DTN, Globus+
  - Fast data transfer

Data Transfer (globus.org)

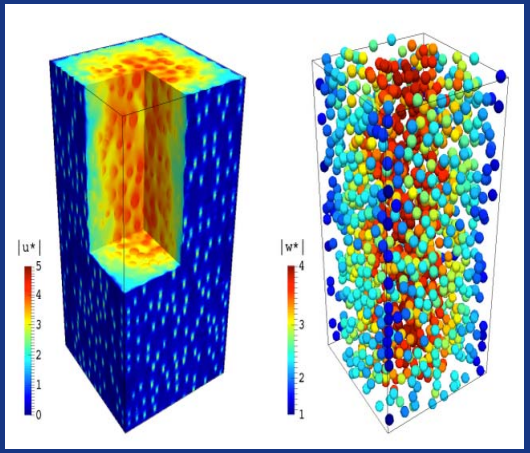


MARCC

JOHNS HOPKINS UNIVERSITY

Showcase

- Prosperetti (1.7M core-hours)
- CFD using the Physalis method (Prosperetti et. al. 2005), which allows direct numerical integration of fluid flows with suspended particles



MARCC

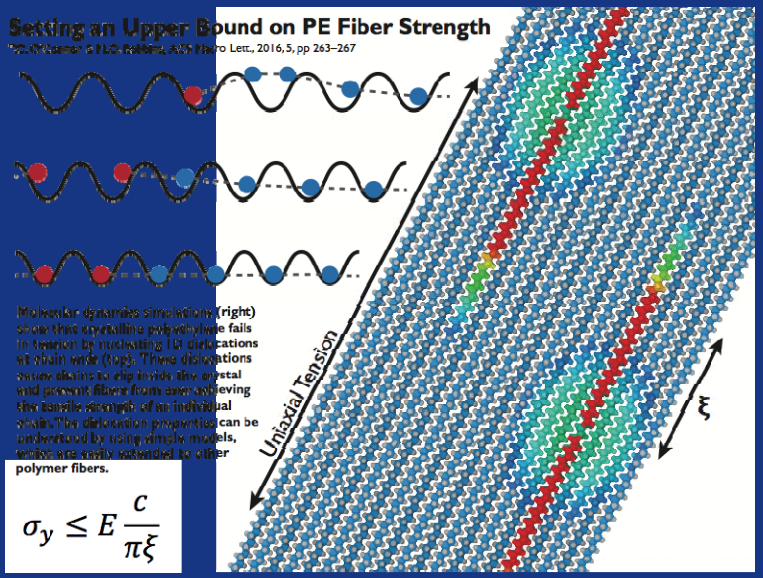
Showcase

JOHNS HOPKINS UNIVERSITY

- Prof. Mark Robbins (13.7M core-hours)
- Ken Hass Outstanding Student Paper Award, "Chain Ends and the Ultimate Tensile Strength of Polyethylene Fibers", APS March meeting Baltimore, 2016.

Setting an Upper Bound on PE Fiber Strength

Lett., 2016, 5, pp 263–267



Molecular dynamics simulations (right) show that crystalline polyethylene fails in tension by nucleating 1D dislocations at chain ends (top). These dislocations cause chains to slip inside the crystal and prevent fibers from ever achieving the tensile strength of an individual chain. The dislocation properties can be understood by using simple models, which are easily extended to other polymer fibers.

$$\sigma_y \leq E\frac{c}{\pi\xi}$$

Data Subject to Restrictions

- Landscape :
  - Traditionally HPC for Science & Engineering == open data
  - Security (system and file integrity)
    - CPU cycles, RAM, Storage
    - Some de-identified data (approval)
- Regulatory Compliance
  - New, unknown concept
  - Additional security, RISK control, single user?
  - Motivation

Reality

- Clinical Research Computing -> confined to Med school small servers
- Public Health a bit less
- Computer Science, Brain Science & BME steady increment
- Need for compute power, storage (usually at HPC)
- With dbGaP data and ePHI data there is a need for HIPAA compliance within HPC Facilities

What is at Stake

- New HIPAA Omnibus Rule (2013)
- Random HIPAA audits (~2014)

- End to end security -> customer end (data generation) -> Network -> data analysis -> data disposal

- Monitor entire research workflow

Standard Data Life-Cycle

Pre-Grant: » Preliminary research » Cyberinfrastructure design » Data Management	Proposal: » Preparation » Budget » Submission/award
Execution: » Data Acquisition » Data Analysis » Data Management » Data Sharing	Post-Grant: » Publishing » Data Archival » Data Disposal » Data Sharing

Confidential-Data Life-Cycle

Pre-Grant: » Preliminary research » *Cyberinfrastructure design* » *Data Management*	Proposal: » Preparation » Budget » *IRB Process*
Execution: » *Data Acquisition* » *Data Analysis* » *Data Management* » *Data Sharing*	Post-Grant: » *Publishing* » *Data Archival* » *Data Disposal* » *Data Sharing*

HIPAA Security Rule

- Requires Administrative, physical and technical safeguards
  - Ensure confidentiality, integrity and availability of all ePHI data
  - Identify and Protect against reasonably identified threats to the security or integrity of the data
  - Protect against reasonable anticipated, improper use or disclosures
  - Ensure compliance by the workforce
  - Provide means for managing risk on an ongoing fashion (process)

- Security Breach Notification laws

Secure Research Environment

- Handle data subject to restrictions, mainly HIPAA
- Based on VMs and VLANs
  - Security, network fragmentation, isolation
- 80% of documentation by MARCC 20% end user
- Based on full cost recovery (business model)
- Needs IRB or higher endorsement

Secure Research Environment (II)

- Base system:
  - VM with 4 cores, 8 GB RAM, 1 TB
- Flexible (per request more cores more RAM)
- Storage
  - Add storage as needed.
- Needs testing . . .
- Capability to add servers & Storage
- Partnership/Collaboration MARCC-Research group

Status

- Availability Early 2017
- Working on cost recovery model/SLA
- Training modules
- Identify best —secure—location
- Future — FISMA Compliance

Thanks

- marcc-help@marcc.jhu.edu
- http://marcc.jhu.edu