Using a Computing Cluster

Dr. Dirk Colbry
Institute for Cyber Enabled Research
Director, High Performance Computing Center
Michigan State University

MICHIGAN STATE UNIVERSITY

ICER

Buzz-Words

- Compute Cluster
- Supercomputers
- Cluster
- High Performance Computing
- High Throughput Computing
- Cyber-Infrastructure
- Advanced Computing Hardware

MICHIGAN STATE UNIVERSITY

ICER

Agenda

- **What is a compute cluster?**
- How do you use a compute cluster
  - Navigating, Developing and testing
  - Accessing installed software
  - Submitting jobs to Scheduler

MICHIGAN STATE UNIVERSITY

ICER

Example: XSEDE

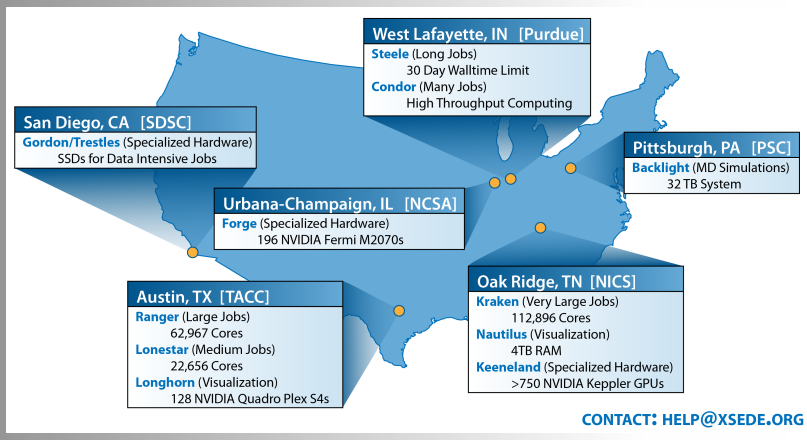
- Formerly TeraGrid
- Funded by the National Science Foundation

"XSEDE is a single virtual system that scientists can use to interactively share computing resources, data and expertise. People around the world use these resources and services — things like supercomputers, collections of data and new tools — to improve our planet."

MICHIGAN STATE UNIVERSITY

ICER

XSEDE Computing Resources

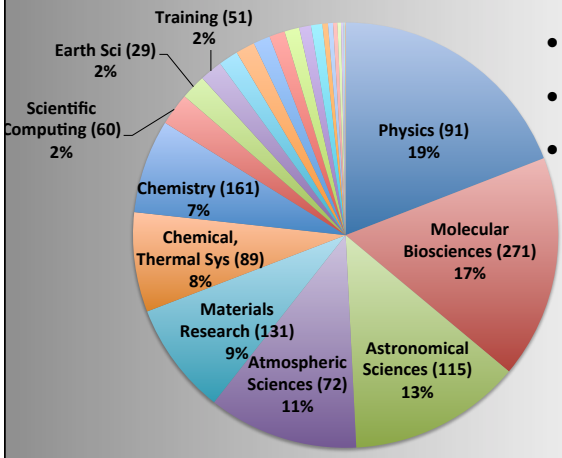


MICHIGAN STATE UNIVERSITY

ICER

Who Uses XSEDE?

- >2 billion cpu-hours allocated
- 1400 allocations
- 350 institutions
- 32 research domains



MICHIGAN STATE UNIVERSITY

ICER

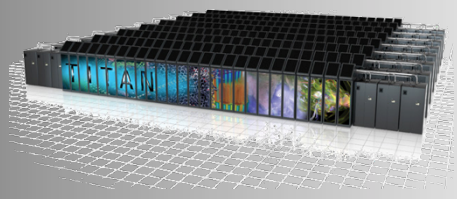
Advanced Computing Hardware

XSEDE
Extreme Science and Engineering Discovery Environment

NASA

Open Science Grid

BLUE WATERS
SUSTAINED PETASCALE COMPUTING



amazon web services™

MICHIGAN STATE UNIVERSITY

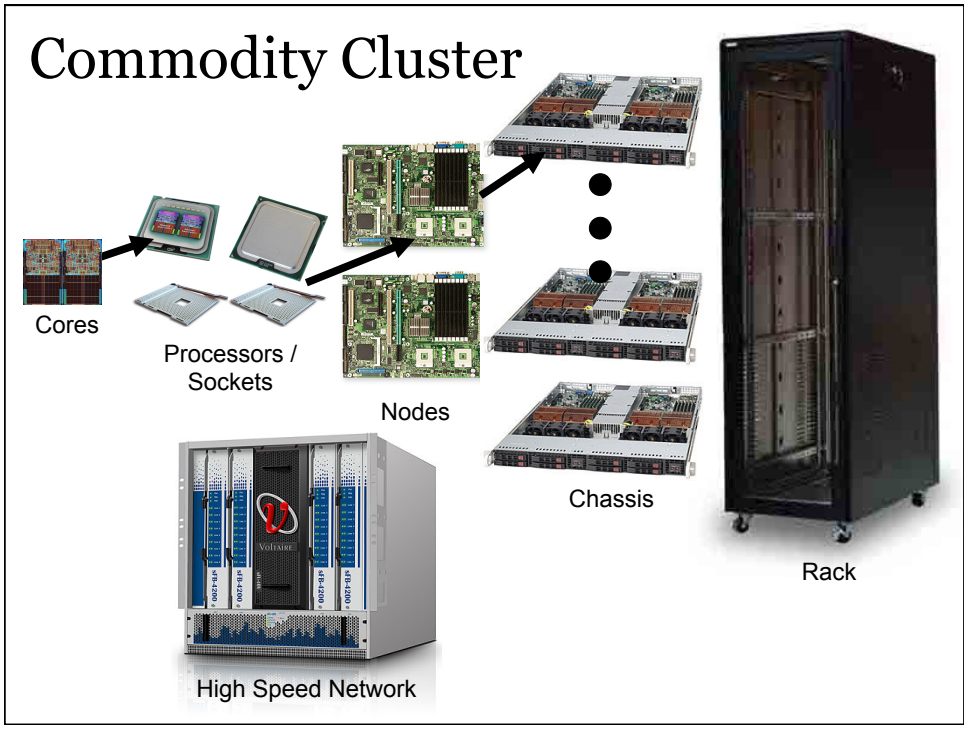
ICER

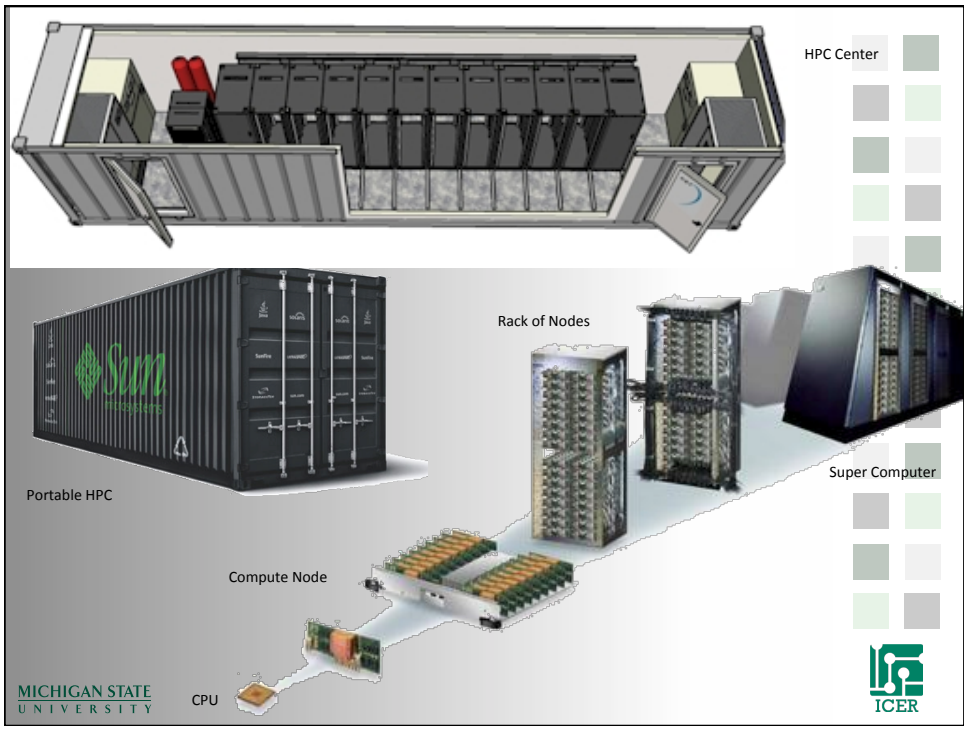
Local Resources

- Lab computers
- Department Resources
- College systems
- Institution systems
  - High Performance Computing Center
- Look for a system near you…

MICHIGAN STATE UNIVERSITY

ICER


Commodity Cluster
Cores
Processors / Sockets
Nodes
Chassis
Rack
High Speed Network


HPC Center
Rack of Nodes
Super Computer
Portable HPC
Compute Node
CPU
MICHIGAN STATE UNIVERSITY
ICER

Large Shared Memory Systems (Fat Nodes)

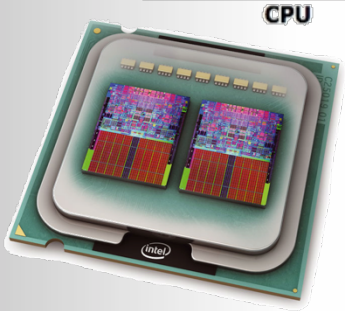
MICHIGAN STATE
UNIVERSITY

ICER

Shared Memory Communication

- Cores on a processor share the same memory
- Many Processors in a box
- OpenMP
- Fat nodes
  - 96 cores
  - 6TB of memory

CPU



MICHIGAN STATE UNIVERSITY

ICER

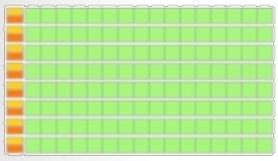
Accelerators

MICHIGAN STATE UNIVERSITY

ICER

GPGPUs

- Cards used to render graphics on a computer
- Hundreds of cores
- Not very smart cores
- But, if you can make your research look like graphics rendering you may be able to run really fast!

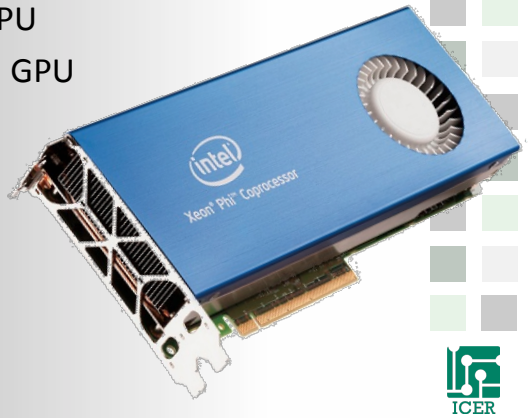


MICHIGAN STATE UNIVERSITY

ICER

Intel Xeon Phi

- Cross between CPU and GPU
- About 60 Pentium I cores
  - Less cores than GPU
  - Easier to use than GPU
    - OpenMP
    - MPI



MICHIGAN STATE UNIVERSITY

ICER

High Throughput HTCondor Cluster

MICHIGAN STATE UNIVERSITY

ICER

MSU HTCondor Cluster

- Runs like a screen saver and Scavenges CPU cycles:
  - Little/No communication between nodes
  - Pleasantly parallel
  - Great for lots-o-small jobs

MICHIGAN STATE UNIVERSITY

ICER

Example: HPCC@MSU

FREE*

- Large Memory Nodes (up to 6TB!)
- GPU Accelerated cluster (K20, M1060)
- PHI Accelerated cluster (5110p)
- Over 600 nodes, 10000 computing cores
- Access to high throughput condor cluster
- 363TB high speed parallel scratch file space
- 50GB replicated file spaces (up to 1TB for free)
- Over 1800 preinstalled user level software packages
- Specialized VMs

MICHIGAN STATE UNIVERSITY

ICER

Which solution is better?

- Depends on what you are doing and your needs:
  - Resources
    - Software
    - Memory
    - File I/O
    - CPU scaling
    - Specialty Hardware
  - Flexibility
  - Price
  - Speed
    - Queue time
    - Run time

MICHIGAN STATE UNIVERSITY

ICER

Agenda

- What is a compute cluster?
- How do you use a compute cluster
  - **Navigating, Developing and testing**
  - Accessing installed software
  - Submitting jobs to Scheduler

MICHIGAN STATE UNIVERSITY

ICER

Exercise: Connect to HPCC

- Step 1: ssh to gateway.hpcc.msu.edu

- Step 2: ssh into a dev node (developer node)

```
> ssh dev-intel10
```

MICHIGAN STATE UNIVERSITY

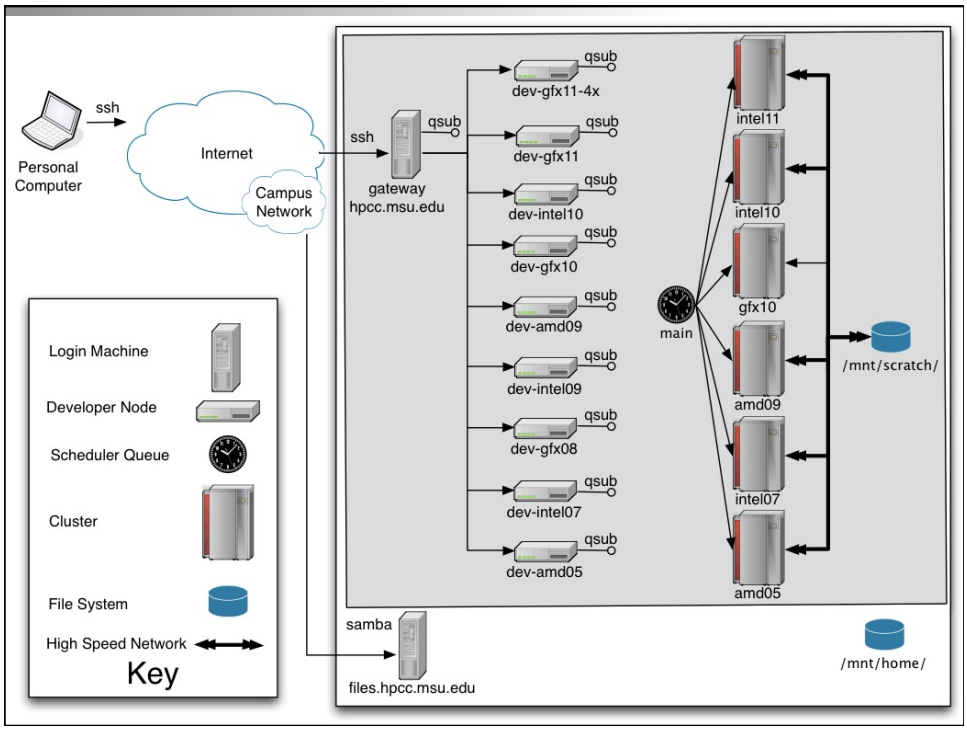
ICER

Key Difference 1

- Gateway vs.
  Head node /Development node vs.
  compute node

MICHIGAN STATE UNIVERSITY

ICER



Running Jobs on the HPC

- Submission scripts are used to run jobs on the cluster
- The developer (dev) nodes are used to compile, test and debug programs

MICHIGAN STATE UNIVERSITY

ICER

Advantages of running Interactively

- You do not need to write a submission script
- You do not need to wait in the queue
- You can provide input to and get feedback from your programs as they are running

MICHIGAN STATE UNIVERSITY

ICER

Disadvantages of running Interactively

- All the resources on developer nodes are shared between all users.
- Any single process is limited to 2 hours of cpu time. If a process runs longer than 2 hours it will be killed.
- Programs that overutilize the resources on a developer node (preventing other to use the system) can be killed without warning.

MICHIGAN STATE UNIVERSITY

ICER

Developer Nodes

Name	Cores	Memory	Accelerators	Notes
dev-intel07	8	8GB	-	
dev-gfx08	4	8GB	3 x M1060	Nvidia Graphics Node
dev-intel10	8	24GB	-	
dev-intel14	20	64GB	-	
dev-intel14-phi	20	128GB	2 x Phi	Xeon Phi Node
dev-intel14-k20	20	128GB	2 x K20	Nvidia Graphics Node

MICHIGAN STATE UNIVERSITY

ICER

Agenda

- What is a compute cluster?
- How do you use a compute cluster
  - Navigating, Developing and testing
  - **Accessing installed software**
  - Submitting jobs to Scheduler

MICHIGAN STATE UNIVERSITY

ICER

Available Software

- Center Supported Development Software
  - Intel compilers, openmp, openmpi, mvapich, totalview, mkl, pathscale, gnu, ...
- Center Supported Research Software
  - MATLAB, R, fluent, abaqus, HEEDS, amber, blast, ls-dyna, starp...
- Customer Software
  - gromacs, cmake, cuda, imagemagick, java, openmm, siesta…
  - For a more up to date list, see the documentation wiki:
    - http://wiki.hpcc.msu.edu/

MICHIGAN STATE UNIVERSITY

ICER

Key Difference 2

- Environment management Systems
  - a.k.a. modules, gnu modules, lua modules, etc

MICHIGAN STATE UNIVERSITY

ICER

Module System

- To maximize the different types of software and system configurations that are available to the users, HPCC uses a Module system
- Key Commands
  - **module avail** – show available modules
  - **module list** – list currently loaded modules
  - **module load** modulename – load a module
  - **module unload** modulename – unload a module
  - **module spider keyword** – Search modules for a keyword

MICHIGAN STATE UNIVERSITY

ICER

Exercise – Module

- List loaded modules

  `>module list`
- Show available modules:

  `>module avail`
- Try an example (Shouldn't work):

  `>powertools`

MICHIGAN STATE UNIVERSITY

ICER

Exercise: getexample

- Load a newly available module:

  `>module load powertools`
- Show powertools (should work now):

  `>powertools`
- Run the "getexample" powertool

  `>getexample`
- Download the helloMPI example

  `>getexample mothur_example`

MICHIGAN STATE UNIVERSITY

ICER

Navigating the example

- Change to the helloworld directory:

```
> cd ./mothur_example
> ls -la
> less README
```

MICHIGAN STATE UNIVERSITY

ICER

Agenda

- What is a compute cluster?
- How do you use a compute cluster
  - Navigating, Developing and testing
  - Accessing installed software
  - **Submitting jobs to Scheduler**

MICHIGAN STATE UNIVERSITY

ICER

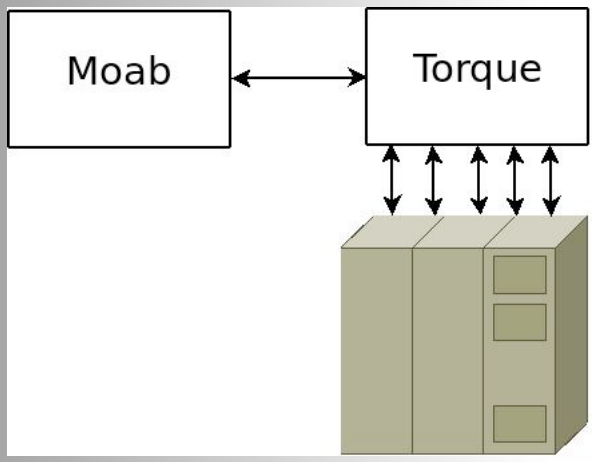
Key Difference 3

- Scheduling system.
- Ex:
  - SLERM
  - PBS Torque / Moab
  - Condor
- Requires a submission script

MICHIGAN STATE UNIVERSITY

ICER

Resource Manager and scheduler



Not First In First Out!!

MICHIGAN STATE UNIVERSITY

ICER

Schedulers vs Resource Managers

- Scheduler
  (Moab)
  - Tracks and assigns
    - Memory
    - CPUs
    - Disk space
    - Software Licenses
    - Power / environment
    - Network
- Resource Manager
  (PBS/Torque)
  - Hold jobs for execution
  - Put the jobs on the nodes
  - Monitor the jobs and nodes

MICHIGAN STATE UNIVERSITY

ICER

Common Commands

- **qsub** <Submission script>
  - Submit a job to the queue
- **qdel** <JOB ID>
  - Delete a job from the queue
- **showq** –u <USERNAME>
  - Show the current job queue
- **checkjob** <JOB ID>
  - Check the status of the current job
- **showstart** –e all <JOB ID>
  - Show the estimated start time of the job

MICHIGAN STATE UNIVERSITY

ICER

Submission Script

1. List of required resources
2. All command line instructions needed to run the computation

MICHIGAN STATE UNIVERSITY

ICER

Typical Submission Script

Shell Comment

Define Shell

```
#!/bin/bash -login
#PBS -l walltime=10:00:00,mem=3Gb,nodes=10:ppn=1
#PBS -j oe

cd ${PBS_O_WORKDIR}

./myprogram -my input arguments

qstat -f ${PBS_JOBID}
```

Resource Requests

Shell Commands

Special Environment Variables

MICHIGAN STATE UNIVERSITY

ICER

Example: Submit a job

- Go to the top helloworld directory

```
>cd ~/mothur_example
```

- Create a simple submission script

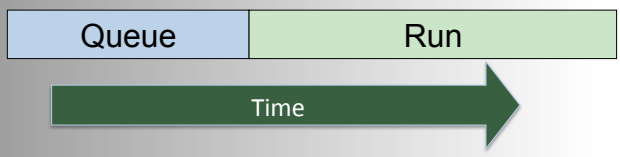
```
>less run_mothur_quality.qsub
```

MICHIGAN STATE UNIVERSITY

ICER

Submitting a job

- qsub –arguments <Submission Script>
  - Returns the job ID. Typically looks like the following:
    - 5945571.cmgr01
- Time to job completion



MICHIGAN STATE UNIVERSITY

ICER

Example: Submit a job, cont.

- Submit the file to the queue

```
>qsub run_mothur_quality.qsub
```

- Record jobid number (######) and wait at most 30 seconds
- Check the status of the queue

```
>showq
```

MICHIGAN STATE UNIVERSITY

ICER

Example: Monitor a job

- Submit the file to the queue:

```
>qstat -f ######
```

- When will a job start:

```
>showstart -e all ######
```

MICHIGAN STATE UNIVERSITY

ICER

Scheduling Priorities

- Jobs that use more resources get higher priority (because these are hard to schedule)
- Smaller jobs are backfilled to fit in the holes created by the bigger jobs
- Eligible jobs acquire more priority as they sit in the queue
- Jobs can be in three basic states:
  - Blocked, eligible or running

MICHIGAN STATE UNIVERSITY

ICER

Cluster Resources

Year	Name	Description	ppn	Memory	Nodes	Total Cores
2007	intel07	Quad-core 2.3GHz Intel Xeon E5345	8	8GB	126	1008
2009	amd09	Sun Fire X4600 (Fat Node) AMD Opteron 8384	32	256GB	3	96
2010	gfx10	NVIDIA CUDA Node (no IB)	8	18GB	32	256
2010	intel10	Intel Xeon E5620 (2.40 GHz)	8	24GB	191	1528
2011	intel11	Intel Xeon 2.66 GHz E7-8837	32	512GB	2	64
			32	1TB	1	32
			64	2TB	2	128
2014	intel14	Intel Xeon E5-2670 v2 (2.6 GHz)	20	64GB	128	2560
			20	256GB	24	480
		2 NVIDIA K20 GPUs	20	128GB	40	800
		2 Xeon Phi 5110P	20	128GB	28	560
	Total				577	7512

MICHIGAN STATE UNIVERSITY

ICER

System Limitations

- Scheduling
  - 5 eligible jobs at a time
  - 520 running jobs
  - 1000 submitted jobs
- Resources
  - 1 week of walltime
  - 520 cores (nodes * ppn)
  - ppn=64
  - 2TB memory on a single core
  - ~200 GB Hard Drive

MICHIGAN STATE UNIVERSITY

ICER

Job completion

- By default the job will automatically generate two files when it completes:
  - Standard Output:
    - Ex: jobname.o5945571
  - Standard Error:
    - Ex: jobname.e5945571
- You can combine these files if you add the join option in your submission script:
  - "#PBS -j oe"
- You can change the output file name
  - #PBS -o /mnt/home/netid/myoutputfile.txt

MICHIGAN STATE UNIVERSITY

ICER

Other Job Properties

- resources (-l)
  - Walltime, memory, nodes, processor, network, etc.
- #PBS –l feature=gpgpu,gbe
- #PBS –l nodes=2:ppn=8:gpu=2
- #PBS –l mem=16gb
- Email address (-M)
  - Ex: #PBS –M colbrydi@msu.edu
- Email Options (-m)
  - Ex: #PBS –m abe

Many others, see the wiki:
http://wiki.hpcc.msu.edu/

MICHIGAN STATE UNIVERSITY

ICER

Summary Clusters vs. Cloud

- Key differences when using:
  - Different node types (gateway, head/ development, compute)
  - Environment management Systems (modules)
  - Batch Submission Script (Resources needed and commands to execute)

MICHIGAN STATE UNIVERSITY

ICER


Discussion
MICHIGAN STATE
UNIVERSITY
ICER