

# Cyberinfrastructure Training: Challenges & Opportunities

*Brian Dobbins*

*NCAR/CGD*

**September 16th, 2022**



# What do we mean by ‘CI training’?

Cyberinfrastructure has a somewhat *fuzzy* definition.

Training students or software engineers:

- Usually focused on learning to use some new kind of *technology*

Instead, I’m going to focus on training *scientific communities*:

- How can we transform and improve the *research process* with CI training?

(Let RSEs develop *capabilities*, and let scientists focus on *science*.)

# NCAR & Cyberinfrastructure

NCAR is a global leader in enabling *earth system science communities*

- We do cutting edge, world-class science
- We also enable others to do science through world-class *cyberinfrastructure*

We're at the *forefront* of new CI; our community looks to us for guidance.

The fundamental challenge:

- Funding for CI *people* is flat while CI *complexity* is growing massively.

# CI: Open Source Models for Science

NCAR develops multiple open source models that enable research:

- CESM, WRF, MPAS-Atmosphere, etc

These are effectively freely available, highly advanced *digital laboratories*.

They are also *inarguably* our most valuable CI asset:

- More than NCAR's scientific output
- More than NCAR's supercomputers

# CESM & Complexity

Model Version	Lines of Code
CSM1.(1996)	70,653
CCSM3 (2004)	247,633
CESM1 (2013)	1,124,312
CESM2 (2018)	1,590,938

The *model* grows in complexity.  
The *community* grows in size.  
The *science* grows in volume.

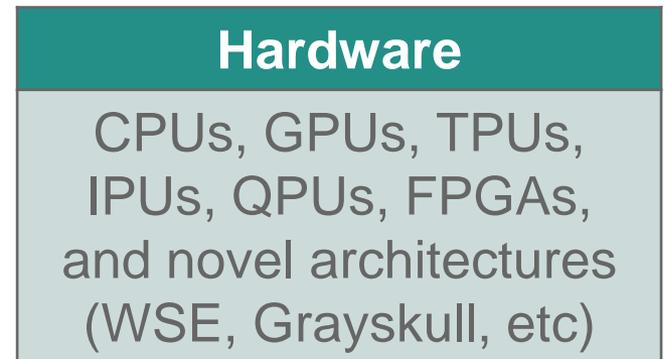
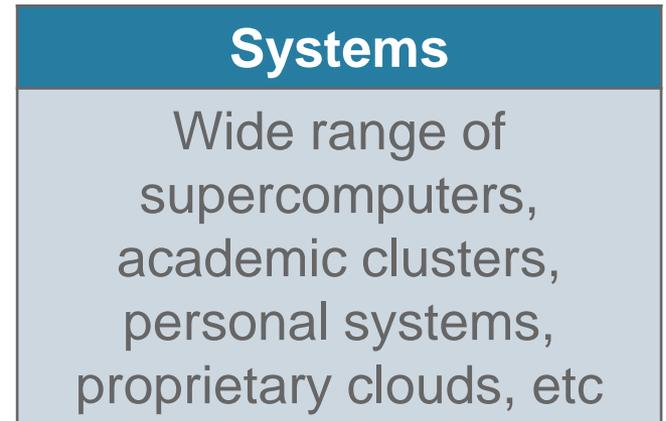
But the funding to *develop & maintain* the model, our critical CI, is flat, *at best*.

# Computing & Complexity

## *Historical CI*



## *Current / Near Future CI*



# The Challenges

The complexity of the CI landscape has grown enormously. The resources we have to support communities have not. The result:

- Staff are spread too thin
- No redundancy – people depart and critical skills are lost
- Competition from private sector for new CI skills
- Reduced opportunities to engage with our community
- Limited time for professional development

Training *communities* in modern CI requires investing in *people* to develop advanced CI capabilities.

# CI Training: Opportunities



# Opportunity: Standardized Platforms

Getting complex CI environments *working* has historically been challenging.

- Often a real impediment to collaborative science!

New technologies like clouds and containers enable us to provide these environments to researchers in *standardized, ready-to-use* form.

- What used to take days, weeks or months now takes *minutes*.
- The end of the ‘tear test’!

Result:

‘Democratized access’ – anyone, anywhere can access sophisticated CI

# Opportunity: Science Training

How often have people here had students come to their facility, learn something, then struggle to replicate it back on their own system?

- Very common across major supercomputing sites!

Leveraging *standardized* platforms, we can develop ‘write once, run anywhere’ training materials for our community.

Result:

Open-source, interactive *curriculum* for science communities.

# Opportunity: Simplified Workflows

Let's rethink even something as simple as *data access* in light of modern CI.

Traditional:

- Google for dataset.
- Find a server with it.
- Download files.
- Transfer to HPC system
- Log in and load up the data in Jupyter

This *works*, but why should it be needed?

```
: dataset = cesmtoolkit.datasets["CESM2/ARISE-SAI-1.5"]
```

Github repo + Python package = Simple, *intelligent* data access!

# Recommendations

For CI *skill* development in students:

- Internships are a *fantastic* opportunity for students and MFs alike.

For *community* CI training:

- Fund 'leadership' RSE teams in specific fields to build *science platforms*.
- Fund the development of curricula / training materials using these platforms.
- Standardized environments make scientists more collaborative and efficient!

Central themes:

RSEs are underfunded for the ever-expanding roles placed on them.

*Additional* funding for them has an outsized effect on science capabilities.

*Extra: Parity between technical and scientific leadership is needed!*

## Extra (For Roland!)

Democratizing access to data – an *excellent* CI project requiring RSE time:

Imagine if scientific data was exchanged as easily as music was during the Napster / filesharing era?

- Concurrent downloads from *local peers*
- Increase total bandwidth by number of peers
- Local caching (incredibly useful!)
- Metrics on access

For NCAR, I've started talking about an implementation called EDEN:  
"Earth Data Exchange Network"

FAIR and simple data access. An RSE project that *improves* science!