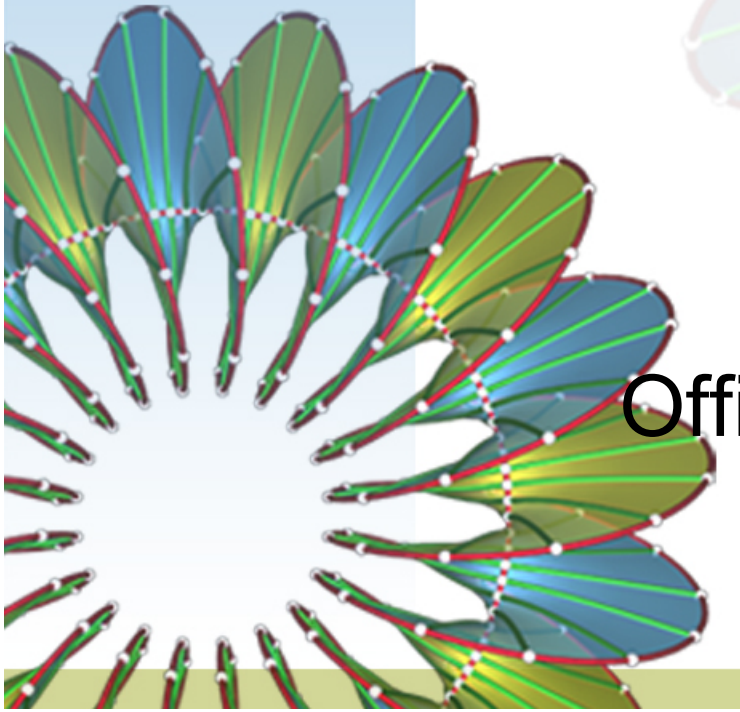
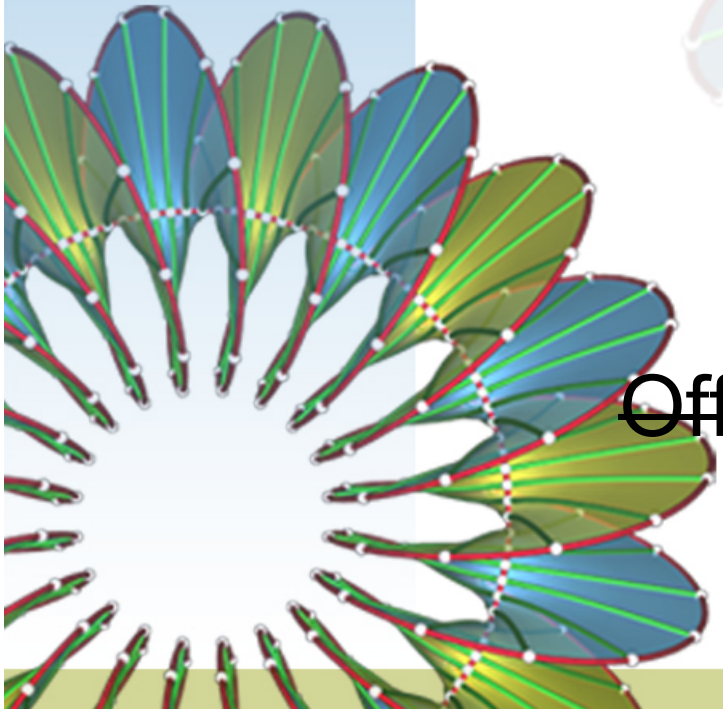


Software as Infrastructure at NSF/OCI

Daniel S. Katz

Program Director,
Office of Cyberinfrastructure (OCI)



Software as Infrastructure at NSF/OCI

Daniel S. Katz

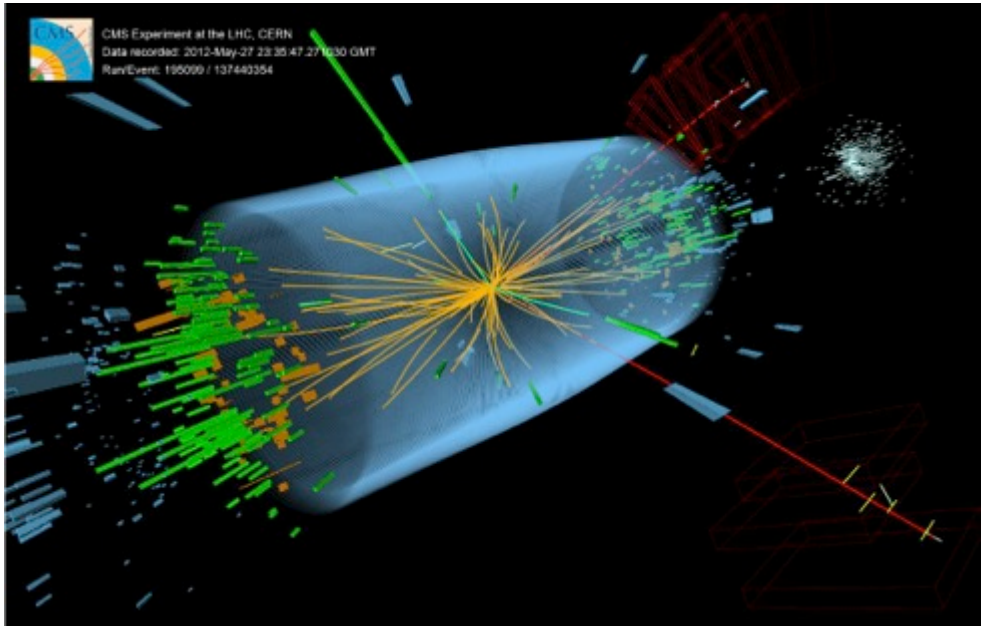
Program Director,

~~Office of Cyberinfrastructure (OCI)~~

Division of Advanced Cyberinfrastructure (ACI)

Big Science and Infrastructure

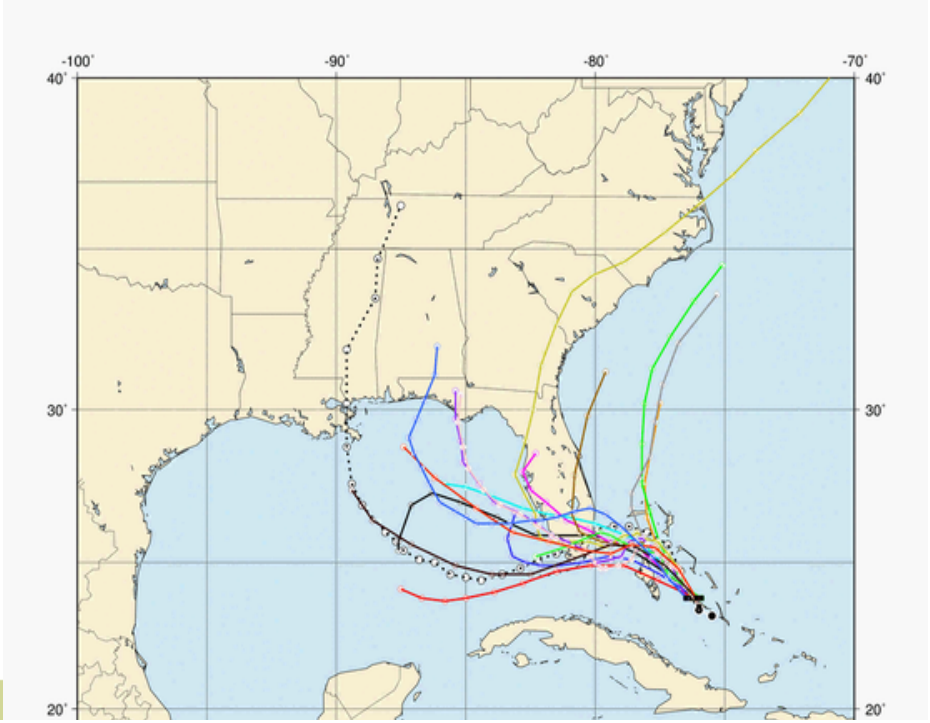
- Higgs* boson discovery announced at CERN July 4, 2012
- Instrument: Large Hadron Collider (LHC)
- Infrastructure
  - Computing Hardware: Worldwide LHC Computing Grid (WLCG): 235,000 cores across 36 countries, including OpenScience Grid (OSG, US), European Grid Infrastructure (EGI, Europe), ...
  - Data: ~20 PB of data created in 2011-2012
  - Software: grid middleware, physics analysis applications, ...
  - Networks
  - Education & Training
- Data generated centrally, moved (~3 PB/week) across multi-tiered infrastructure to be computing upon



See: http://www.isgtw.org/feature/how-grid-computing-helped-cern-hunt-higgs

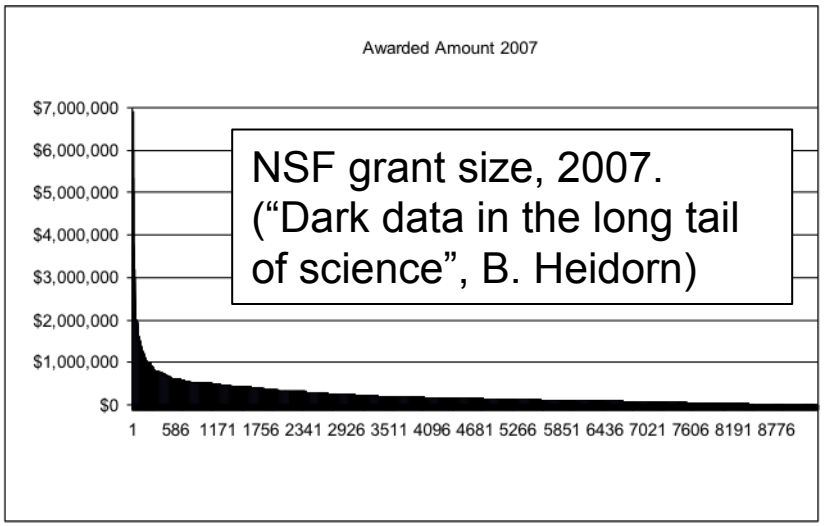
Big Science and Infrastructure

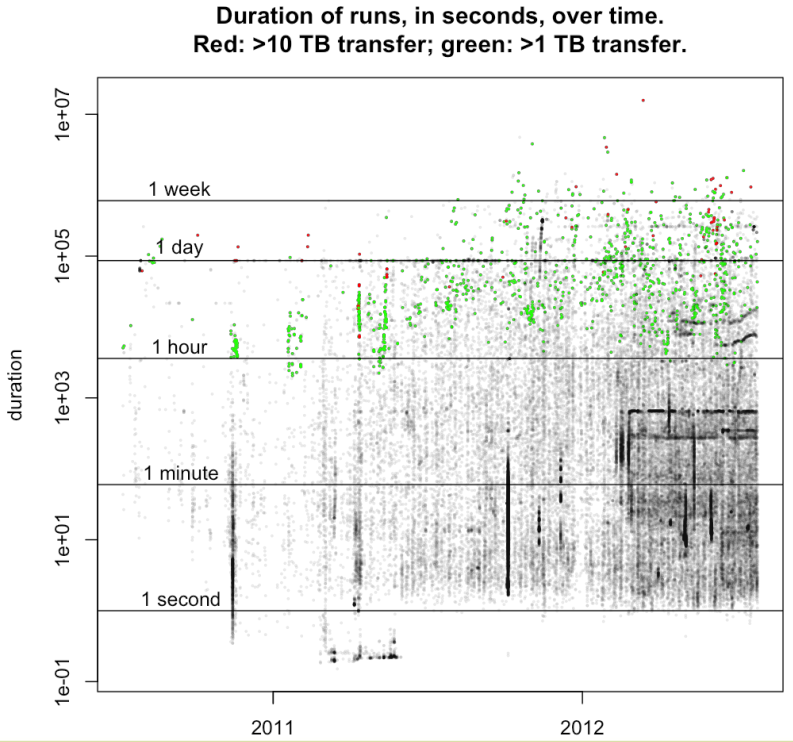
- Hurricanes affect humans
- Multi-physics: atmosphere, ocean, coast, vegetation, soil
  - Sensors and data as inputs
- Humans: what have they built, where are they, what will they do
  - Data and models as inputs
- Infrastructure:
  - Urgent/scheduled processing, workflows
  - Software applications, workflows
  - Networks
  - Decision-support systems, visualization
  - Data storage, interoperability



Long-tail Science and Infrastructure

- Exploding data volumes & powerful simulation methods mean that more researchers need advanced infrastructure
- Such “long-tail” researchers cannot afford expensive expertise and unique infrastructure
- Challenge: Outsource and/or automate time-consuming common processes
  - Tools, e.g., Globus Online and data management
    - Note: much LHC data is moved by Globus GridFTP, e.g., May/ June 2012, >20 PB, >20M files
  - Gateways, e.g., nanoHUB, CIPRES, access to scientific simulation software





Long-tail Science and Infrastructure

- CIPRES Science Gateway for Phylogenetics
  - Study of diversification of life and relationships among living things through time
- Highly used
  - Cited in at least 400 publications, e.g., Nature, PNAS, Cell
  - More than 5000 unique users in 3 years
  - Used routinely in at least 68 undergraduate classes
  - 45% US (including most states), 55% 70 other countries
- Infrastructure
  - Flexible web application
    - A science gateway, uses software and lessons from XSEDE gateways team, e.g., identify management, HPC job control
  - Science software: tree inference and sequence alignment
    - Parallel versions of MrBayes, RAxML, GARLI, BEAST, MAFFT
    - PAUP*, Poy, ClustalW, Contralign, FSA, MUSCLE, ...
  - Data
    - Personal user space for storing results
    - Tools to transfer and view data

Credit: Mark Miller, SDSC

Infrastructure Challenges

- Science
  - Larger teams, more disciplines, more countries
- Data
  - Size, complexity, rates all increasing rapidly
  - Need for interoperability (systems and policies)
- Systems
  - More cores, more architectures (GPUs), more memory hierarchy
  - Changing balances (latency vs bandwidth)
  - Changing limits (power, funds)
  - System architecture and business models changing (clouds)
  - Network capacity growing; increase networks -> increased security
- Software
  - Multiphysics algorithms, frameworks
  - Programing models and abstractions for science, data, and hardware
  - V&V, reproducibility, fault tolerance
- People
  - Education and training
  - Career paths
  - Credit and attribution

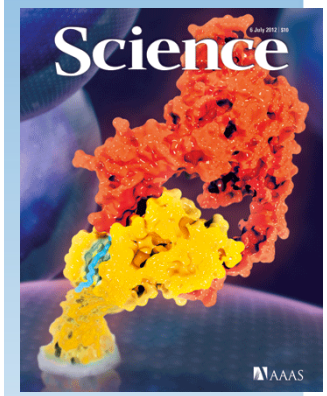
Cyberinfrastructure (e-Research)

- *"Cyberinfrastructure consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible."*

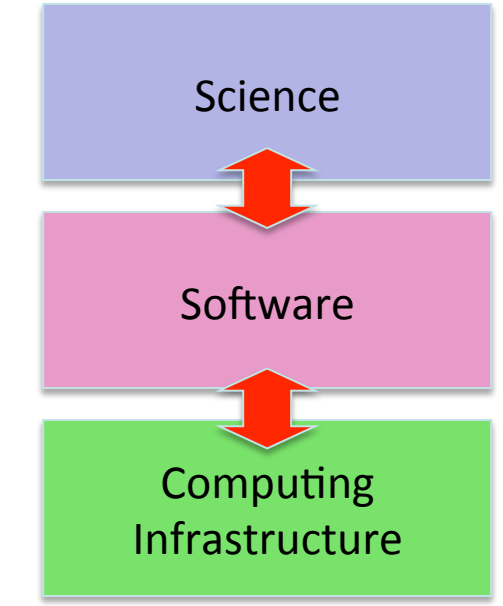
  -- Craig Stewart

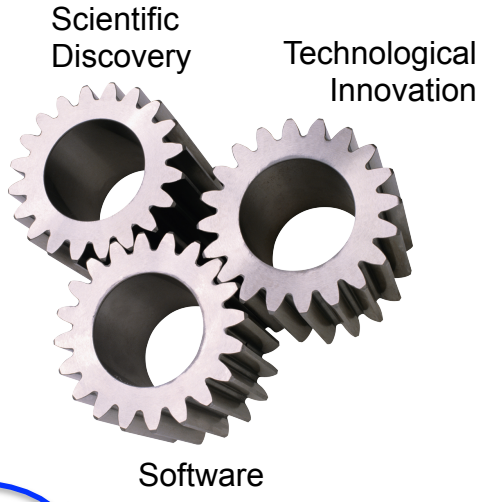
- Infrastructure elements:
  - parts of an infrastructure,
  - developed by individuals and groups,
  - international,
  - developed for a purpose,
  - used by a community

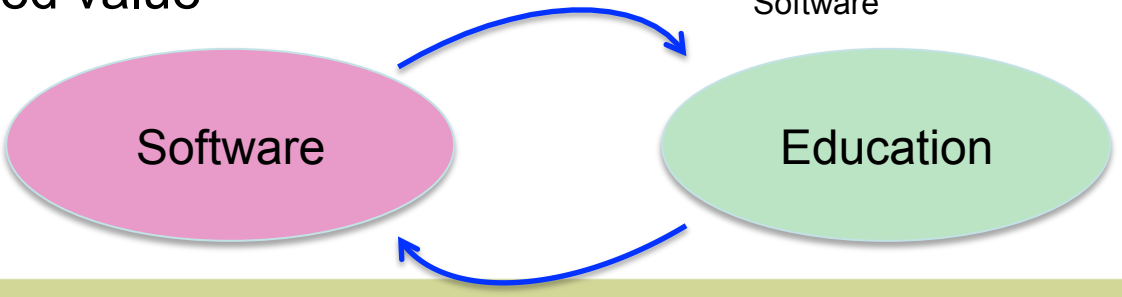
Software is Infrastructure



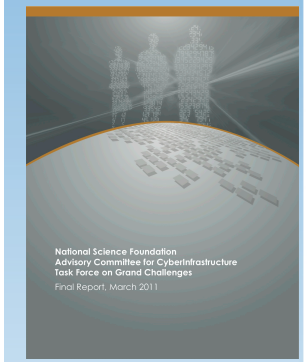
- Software essential for the bulk of science
  - About half the papers in recent issues of Science were software-intensive projects
  - Research becoming dependent upon advances in software
  - Significant software development being conducted across NSF: NEON, OOI, NEES, NCN, iPlant, etc
- Wide range of software types: system, applications, modeling, gateways, analysis, algorithms, middleware, libraries
- Development, production and maintenance are people intensive
- Software life-times are long compared to hardware
- Under-appreciated value

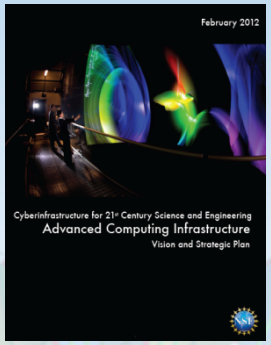






Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21)

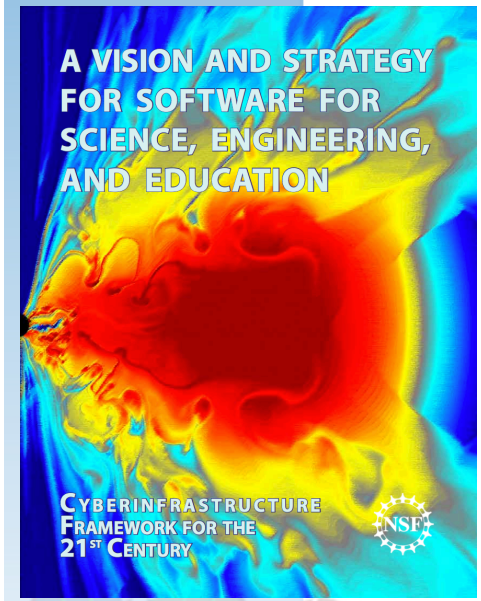




- Cross-NSF portfolio of activities to provide integrated cyber resources that will enable new multidisciplinary research opportunities in all science and engineering fields by leveraging ongoing investments and using common approaches and components (http://www.nsf.gov/cif21)

- ACCI task force reports (http://www.nsf.gov/od/oci/taskforces/index.jsp)
  - Campus Bridging, Cyberlearning & Workforce Development, Data & Visualization, Grand Challenges, HPC, Software for Science & Engineering
  - Included recommendation for NSF-wide CDS&E program
- Vision and Strategy Reports
  - ACI - http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf12051
  - Software - http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf12113
  - Data - http://www.nsf.gov/od/oci/cif21/DataVision2012.pdf
- Implementation
  - Implementation of Software Vision http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504817

Software Vision



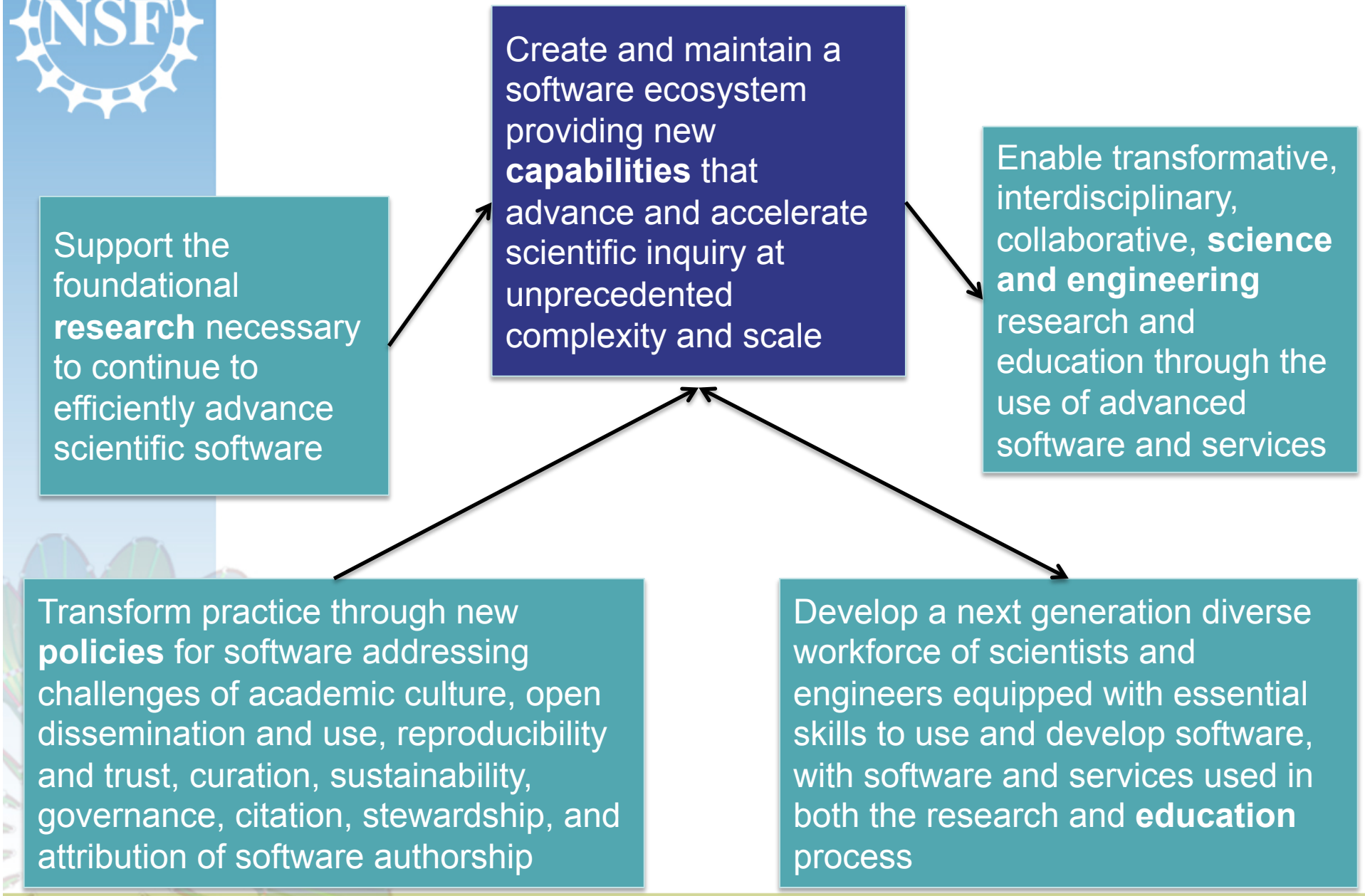
NSF will take a leadership role in **providing software as enabling infrastructure** for science and engineering research and education, and in **promoting software** as a principal component of its comprehensive CIF21 vision

...

Reducing the complexity of software will be a unifying theme across the CIF21 vision, **advancing** both the **use and development of new software** and **promoting the ubiquitous integration of scientific software across all disciplines**, in education, and in industry

- A Vision and Strategy for Software for Science, Engineering, and Education – NSF 12-113

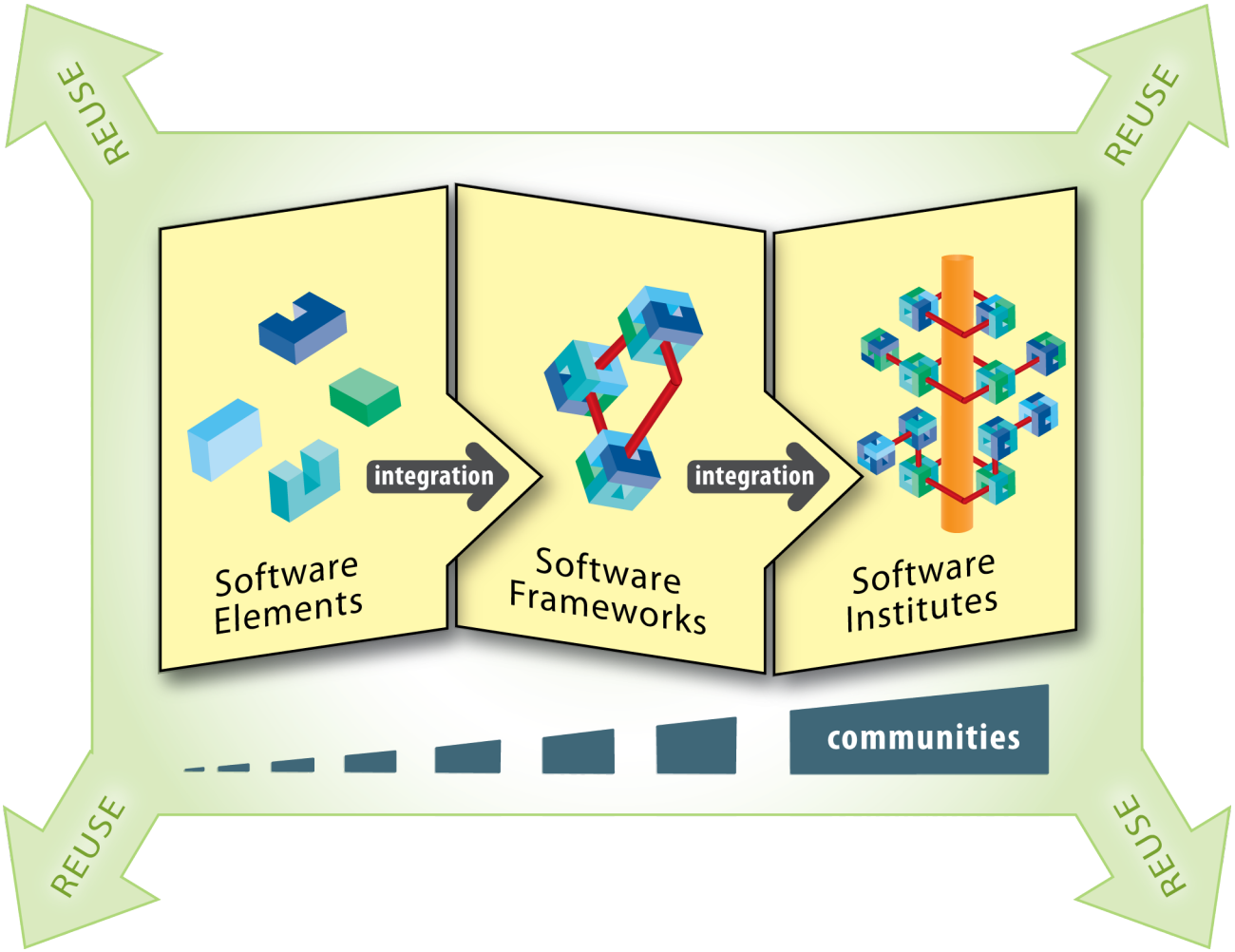
Infrastructure Role & Lifecycle



ACI Software Cluster Programs

- Exploiting Parallelism and Scalability (XPS)
  - New CISE & OCI program for foundational groundbreaking research leading to a new era of parallel (and distributed) computing
  - Issued in Oct., proposals submitted in Feb.
- Computational and Data-Enabled Science & Engineering (CDS&E)
  - Virtual program (ENG, MPS, OCI) for science-specific proofing of algorithms and codes
  - Identify and capitalize on opportunities for major scientific and engineering breakthroughs through new computational and data analysis approaches
- Software Infrastructure for Sustained Innovation (SI^2)
  - Transform innovations in research and education into sustained software resources that are an integral part of the cyberinfrastructure
  - Develop and maintain sustainable software infrastructure that can enhance productivity and accelerate innovation in science and engineering

Software Infrastructure Projects



SI^2 Software Activities

- Elements (SSE) & Frameworks (SSI)
  - Past general solicitations, with most of NSF (BIO, CISE, EHR, ENG, MPS, SBE): NSF 10-551 (2011), NSF 11-539 (2012)
    - About 27 SSE and 20 SSI projects (19 SSE & 13 SSI in FY12)
  - Current focused solicitation, with MPS/CHE and EPSRC: US/UK collaborations in computational chemistry, NSF 12-576 (2012)
    - Will fund 4 awards from 18 proposals
  - Solicitation open (NSF 13-525)
- Institutes (S2I2)
  - Solicitation for conceptualization awards, NSF 11-589 (2012)
    - 13 projects (co-funded with BIO, CISE, ENG, MPS)
  - Second solicitation for 3-5 more S2I2s (NSF 13-511)
  - Full institute solicitation in late FY14
- US/China DCL (with CISE/CNS, loosely with NSFC)
  - NSF 12-096: will make decisions soon on small set of initial projects
  - Included in fuure SSE&SSI solicitation
- See http://bit.ly/sw-ci for current projects

SI^2 Solicitation and Decision Process

- Cross-NSF software working group with members from all directorates
- Determined how SI^2 fits with other NSF programs that support software
  - See: Implementation of NSF Software Vision - http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504817
- Discusses solicitations, determines who will participate in each
- Discusses and participates in review process
- Work together to fund worthy proposals

SI^2 Solicitation and Decision Process

- Proposal reviews well -> my role becomes matchmaking
  - I want to find program officers with funds, and convince them that they should spend their funds on the proposal
- Unidisciplinary project (e.g. bioinformatics app)
  - Work with single program officer, either likes the proposal or not
- Multidisciplinary project (e.g., molecular dynamics)
  - Work with multiple program officers, ...
- Onmidisciplinary project (e.g. http, math library)
  - Try to work with all program officers, often am told “it’s your responsibility”
- In all cases, need to forecast impact
  - Past performance does predict future results

Measuring Impact – Scenarios

1. Developer of open source physics simulation
   - Possible metrics
     - How many downloads? (easiest to measure, least value)
     - How many contributors?
     - How many uses?
     - How many papers cite it?
     - How many papers that cite it are cited? (hardest to measure, most value)
2. Developer of open source math library
   - Possible metrics are similar, but citations are less likely
   - What if users don't download it?
     - It's part of a distro
     - It's pre-installed (and optimized) on an HPC system
     - It's part of a cloud image
     - It's a service

Vision for Metrics & Citation, part 1

- Products (software, paper, data set) are registered
  - Credit map (weighted list of contributors—people, products, etc.) is an input
  - DOI is an output
  - Leads to transitive credit
    - E.g., paper 1 provides 25% credit to software A, and software A provides 10% credit to library X -> library X gets 2.5% credit for paper 1
    - Helps developer – “my tools are widely used, give me tenure” or “NSF should fund my tool maintenance”
  - Issues:
    - Social: Trust in person who registers a product
      - This seems to work for papers today (without weights) for both author lists and for citations
      - Do weights require more than human memory?
    - Technological: Registration system
      - Where is it/them, what are interfaces, how do they work together?

Vision for Metrics & Citation, part 2

- Product usage is recorded
  - Where?
    - Both the developer and user want to track usage
    - Privacy issues? (legal, competitive, ...)
    - Via a phone home mechanism?
  - What does “using” a data set mean? And how could trigger a usage record
  - Can general code be developed for this, to be incorporated in software packages?
- Ties to provenance
- With user input, tie later products to usage
  - User may not know science outcome when using tool
  - After science outcome is known, may be hard to determine which product usages were involved

Vision for Metrics & Citation, thoughts

- Can this be done incrementally?
- Lack of credit is a larger problem than often perceived
  - Lack of credit is a disincentive for sharing software and data
  - Providing credit would both remove disincentive as well as adding incentive
  - See Lewin's principal of force field analysis (1943)
- For commercial tools, credit is tracked by $
  - But this doesn't help understand what tools were used for what outcomes
  - Does this encourage collaboration?
- Could a more economic model be used?
  - NSF gives tokens are part of science grants, users distribute tokens while/after using tools

Software Questions: Sustainability

- My definition as a program officer:
  - How will you support your software without me continuing to pay for it?
- What does support mean?
  - Can I build and run it on my current system?
    - Adapt to changing underlying hardware/software
  - Do I understand what it does?
    - Documentation, training
  - Does it do what it does correctly?
    - Bug tracking and updates
    - Verification and validation
  - Does it do what I want?
    - Requirement tracking and updates
  - Is it changing?
    - Heritage (legacy) vs. developing software
- How can Apache help?

Software Questions: Governance

- Why do we care?
  - Governance tells users and contributors how the project makes decisions, how they can be involved
- What are the issues?
  - Community: Users? Developers? Both?
  - Models: dictatorship (Linux kernel), meritocracy (Apache), other?
  - Tie to development models: cathedral, bazaar
- How can Apache help?
  - Study how these work in smaller specialized projects?

Other Questions for Apache

- Does the Apache Way work for science?
  - Or just for underlying tools that are useful both for science and other applications?
- How many users/developers are needed for success?
- Incubator model
  - Can it be used as is for general science software?
  - Or forked and modified?
- Open Source for understanding (available) vs Open Source for reuse/development (changeable)?

General Software Questions

- Software that is intended to be infrastructure has challenges
  - Unlike in business, more users means more work
  - The last 20% takes 80% of the effort
  - What can NSF do to make these things easier?
- What fraction of funds should be spent of support of existing infrastructure vs. development of new infrastructure?
- How do we decide when to stop supporting a software element?
- How do we encourage reuse and discourage duplication?
- How do we more effectively support career paths for software developers (with universities, labs, etc.)

What Can You Do?

- Look at the current set of SI^2 software and institutes, and get involved with one
  - http://bit.ly/sw-ci
- Tell me what we should be doing differently
  - Here or email: dkatz@nsf.gov