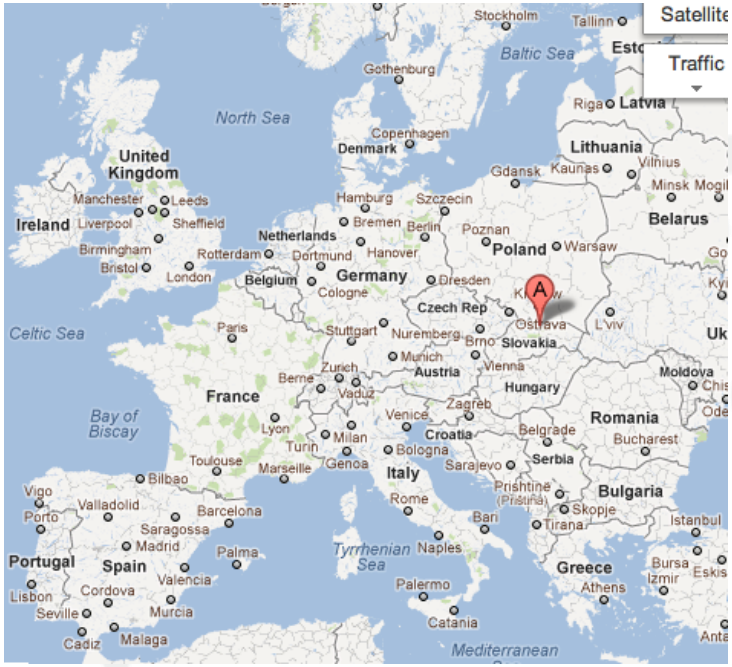
Computation Institute

Grid Computing: the Next Decade

Daniel S. Katz
Senior Fellow
University of Chicago & Argonne National Laboratory
Representing the workshop organizers: Krzysztof Kurowski, Jarek Nabrzyski, Andre Merzky

Meeting

- Grid Computing: the Next Decade
- 4th – 6th January 2012, Zakopane, Poland
- http://www.gridlab.org/Meetings/Zakopane2012
- 10th year anniversary of the FP5 GridLab project (12 EU, 2 industry, 3 US partners)
- Main focus: Think about new levels of cooperation, integration of cyberinfrastructure efforts needed to support global science in the next decade.
  - Who are the stakeholders and how should we engage them?
  - What are science needs, particularly in regard to data challenges and new grand challenge communities?
  - What are global issues for cooperation and sharing?
  - Should there be a "blueprint" for CI for global science?
  - Learn from previous experiences in last decade



Participants

- Organizers
  - Krzysztof Kurowski (PSNC), Jarek Nabrzyski (ND), Andre Merzky (LSU), Daniel S. Katz (UC/ANL)
- Agencies
  - Gabrielle Allen (NSF/OCI), Kostas Glinos (EC), Ed Seidel (NSF/MPS)
- 50 attendees
  - Infrastructure/tool providers & application community participation from US/EC/Asia (some US travel funded by NSF)
    - US: Ewa Deelman (USC/ISI), Rion Dooley, Ian Fisk (CMS), Ian Foster (Globus), Geoffrey Fox (FutureGrid), David Hart (NCAR), Shantenu Jha (SAGA), Daniel Katz (UC/ANL), Miron Livny (OSG), Maciej Malawski (ND), Jarek Nabrzyski (remote), Judy Qiu (IU), Karolina Sarnowska-Upton (UVa), Michela Taufer (UDel), Doug Thain (ND), John Towns (XSEDE), Von Welch (IU)
    - EC: Cees De Laat (SARA), Morris Riedel (PRACE), Peter Coveney (UCL), Steven Newhouse (EGI), etc.
    - Satoshi Sekiguchi (NAREGI, Japan), Hai Zhuge (CAS, China)

Dynamic Meeting Format

- Visionary keynotes
  - Ian Foster, Computation Institute: “Scientific Computing in 2020: Grids, Clouds, Skies, ...”
  - Hai Zhuge, Chinese Academy of Science: “Knowledge Grid”
  - Miron Livny, UWisconsin, USA: “What will it take to keep us accountable for our promises?”
  - Peter Coveney & Nour Shublaq, UCL, UK: “Distributed Computing: The Whole Truth and Nothing but the Truth”
  - Anthony Tyson, UC Davis: “Enabling Exascale Exploration and Discovery”
  - Alexander Szalay, Johns Hopkins University: “Data Driven Discovery in Science”

Breakout Groups

- Much of the meeting organized into breakout groups
  - Big science
  - Long tail of science
  - New grand challenge communities
- Then a group-wide consensus on how to move forward

Initial discussions - Requirements

- The three breakout groups were asked to address:
  1. What are the science requirements for particular community? (big science, long tail science, grand challenge communities)
  2. How can we define a “blueprint” or “architecture” to provide an enabling CI for global science in the next decade?
- They also tried to define their own identity, in order to get on a common footing and to start the discussion.

Big Science

- Defined as
  - Large project budget
  - Large collaboration in terms of number of people
  - Extremely large amount of data generated
  - Large complexity of instruments and the large complexity of the required oversight
  - Long timescales
  - All of the above
- A common property though is, if compared to Small Sciences / Long Tail, that there is generally a more structured approach to work with end user requirements (simply because nothing else scales)
- Also, it implies at least some long term planning

Big Science Requirements and Planning

- Generally an effort to collect requirements before project starts
  - With a series of workshops, etc.
- And ongoing efforts to stay connected to user community while evolving the CI stack
- But there's often no long-term funding commitment, just a series of 3-5 year "projects"
  - So planning for CI to support long-term "big" science efforts is hard

Long Tail Science

- Defined as
  - Small collaborations, often ad hoc or spontaneous
  - Parts of a larger community (e.g., Biology)
  - Multidisciplinary
  - Small project size (e.g., “98% of NSF grants are $1m or less”)
  - Minor planned role of information technology
  - Democratized access (citizen science)
- Also called small science, not pejorative
- While in large projects, efforts to gather community requirements can be supported, in small science, no one is funded to do this

Small Science Requirements and Planning

- Could be done by going through a bunch of examples, pulling out requirements, and then looking to see which are common
- Could also be done by looking at what scientists are doing and how they are doing it, then looking at what they say they want to do and what they are missing that would enable them to do these things
  - E.g., for a sequencing user, for example, these include: data movement; metadata generation; data storage; data search; provenance; sharing; tagging; and analysis.
- Similar studies have been done previously, and rather than duplicating them, we should look at their outputs. For example, two completing teams did this for the NSF XD solicitation, and then merged their gathered requirements into one set:
  - XSEDE Requirements, to be published at http://www.xsede.org/publications

Grand Challenge Communities

- Defined as the next generation of "Grand Challenge Projects" addressing global scientific problems that are too large and diverse even for a consortium of research groups
- Example problems include
  - Modeling and understanding of gamma ray bursts requiring many different disciplines (gravitational physics, astrophysics, chemistry, mathematics, computer science)
  - Particle physics, gravitational science, or astronomy motivated by large instruments such as the Large Hadron Collider, LIGO/VIRGO, LSST, etc.
  - Understanding the human brain

Grand Challenge Requirements and Planning (1)

- Focus on understanding what is needed to support these communities, not necessarily specific to the grand challenges themselves, but the incremental needs to support the research communities
- One particular focus was around improving the access to existing research, data and instruments, for example improving access for students
- Aspects related to access
  1. Technical infrastructure
  2. Policies such as intellectual property management and OpenAccess
  3. Social and cultural interactions and understanding how communities cooperate
  4. Regulations for data, such as data privacy
- Universities and institutions were seen to have a responsibility to play a role
  - E.g., the PSNC data services team assists affiliated universities with services such as optical networks, identity management, videoconference/HDTV and broadcast capabilities, cloud capabilities including license centralization, data archive and back up

Grand Challenge Requirements and Planning (2)

- For research projects, needs are:
  - Policies that encourage data to be made available and define validation standards
  - Places to put the data where it can be linked to publications, as well as powerful search engines to find the most relevant and useful data
  - Systems supporting the activity of managing the use of data from its point of creation to ensure it is available for discovery and re-use in the future
- Research data lifecycle management systems are hard to implement, especially for grand challenge communities that have a long history of operations, and never addressed this issue
  - E.g., HEP community has just started some planning activities oriented towards developing data preservation solutions.
- Summary:
  1. Support the progressive advancement of scientific work
  2. Management and incentivation of social interactions
  3. Technical data storage and identification
  4. Access to consistent computing resources and core data services
  5. Funding model and governance for global research communities
  6. Training of students

Second discussion: Process

- The breakout groups were asked to discuss
  1. How would we develop a (minimal) high level blueprint/framework/conceptual architecture or set of processes (or is there a better word?) to organize and coordinate the development and support of cyberinfrastructure, e.g. could expect that this would include
     - minimal security assurances, identity management
     - data sharing policies
     - collaborative software development
     - campus bridging to international infrastructures
     - governance mechanisms
     - continued innovation, as illustrated by the rapid progress of commercial offerings
     - reuse and best practices
  2. How would these processes aid in activities such as
     - Sustainability
     - international cooperation
     - any others?
  3. How to turn this “blueprint” into a set of actionable processes?

Big Science

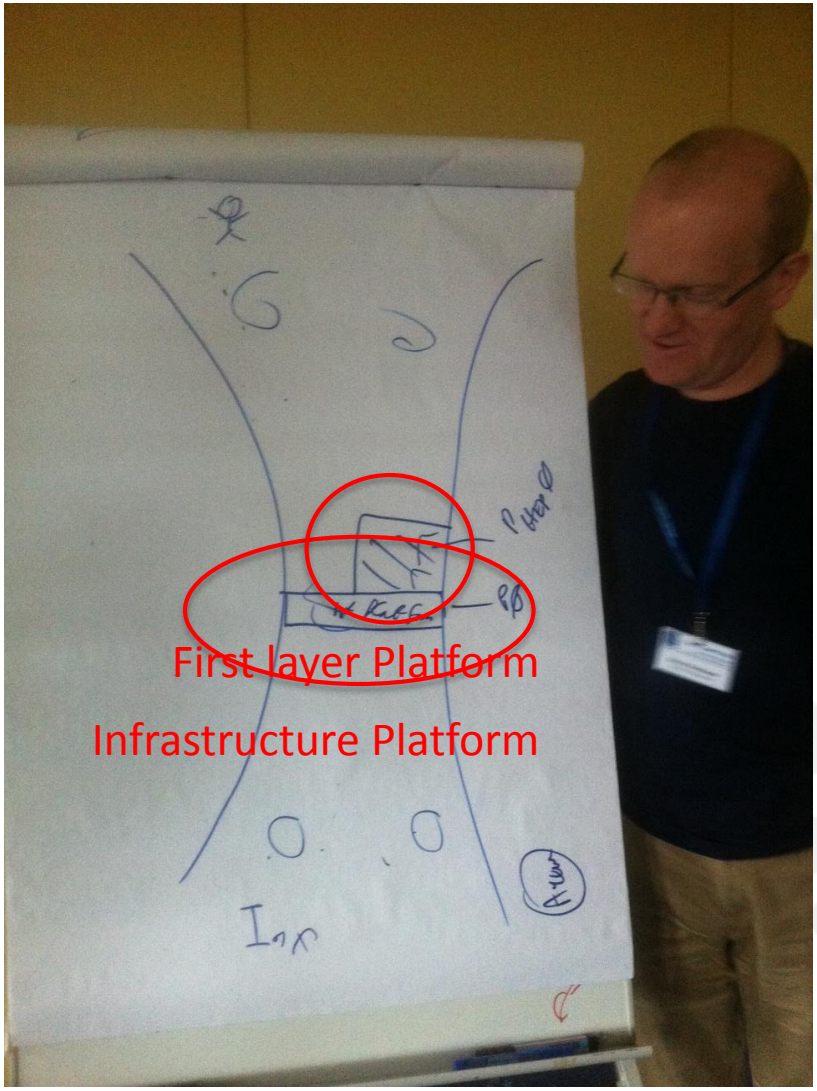
- Strong (but not universal) agreement that adoption of existing solutions is preferable to implementing new solutions - but that does not go well in the current funding structure
  - Funding agencies fund solutions, not processes (Earthcube is an exception here)
- Large international CIs (EGI, PRACE, NAREGI, XSEDE) may be able to coordinate on that level, but user communities will likely only participate when there is an obvious benefit, not before
- Ideally, the blueprint process will support adoption of existing solutions, and thus greatly increase coherence and sustainability of the CI stack
- While it is hard to distinguish needs from desires, the blueprint should focus on needs in order to be manageable and efficient
- An open process for a global CI blueprint can obviously get unwieldy, and multiple funding agencies complicates things further, but a clear demonstration on return-of-investment and a clear process will help to get buy-in
- A blueprint would pose a very significant amount of work, but there is agreement that this effort is worthwhile, and economically beneficial
- Nevertheless, it requires significant commitment from the key stakeholders
- There is consensus that a global blueprint process is needed – even if 'global' is not acceptable for some communities
- While funding agencies are amongst the stakeholders, they should not own the blueprint process, but support it, implicitly and explicitly
- The process would be most efficient if owned by a relatively small group with global input – which is a difficult balance

Small Science

- Two possible processes – standardized interfaces or policies
  - Standardized interfaces tried in OGF, useful but not completely successful
    - Need to have multiple people in projects where the standard would be applied who want to define a standard
    - Standards needs to be developed with consideration of operational issues; people who will implement the standard in working systems need to be in the discussion
- Should the process be top-down or bottom-up?
  - Perhaps a difference in how large-science and small-science sees things
  - Support for ecosystem model, with “market”-based approach, where possible

Small Science: Multi-Level Architecture

- Use cyclic process - start very general, then iterate, getting more specific each cycle (define minimal services/interfaces/platform that one can build on first)
  - Platform needs to track advances in underlying technology if possible
- Infrastructure platform provides common interface to infrastructures
  - Design is iterative
- Layers above (platforms) build on this and provide higher level capabilities
  - Design is iterative
  - E.g., Hadoop, HEP platform
- Some common services also must be provided to all layers
  - E.g., identity management



Small Science: Ecosystem

- Need to decide governance of process - e.g., is blueprint binding?
  - Perhaps this is cyclic – binding later in the process
  - Might need funding agencies to require this to be binding
- Model of app store (tracks usage, reviews, success stories) for components that are sufficiently low-cost, but perhaps need to plan coordinated/shared development/usage of larger activities (perhaps success stories is outside, and points back to app store)
  - Stakeholders need confidence that shared/coordinated software will really appear, do what they want, and will continue to exist
  - Multiple app stores (one for each infrastructure)?
    - Maybe needed today, but not desirable – depends on if we have sufficient standards, or a common infrastructure
    - Could make integrated app store that hides apps that won't work on a particular infrastructure/platform
    - Could perhaps have federated apps stores that look like one, but need single trusted point for metrics of usage.
  - Need to define what's common across infrastructures, and what's not

Small Science – Process Summary

1. Focus on solutions that are relatively simple individually, and which can be built into more by combining multiple of these solutions. On top of such general building blocks, one could imagine additional layers that are less general, perhaps customized for a community, such as a science domain or a technology like Hadoop. Steven Newhouse presented a slide that discussed “Community Services”, which was based on this idea.
2. Use cycles of workshops to discuss elements
   - start with architecture workshop, where infrastructure providers discuss ...
   - also software workshop, e.g. http://sciencesoft.web.cern.ch to discuss shared software repositories
3. In some cases, doing small simple things at large scale can be profound. For example, one could take an analysis that runs on a desktop and scale it out to run 100 to 100,000 times at once. In this way, the small science that is done in the long tail becomes large science.
4. Layers can provide a scaling mechanism, even if they are not simple. Different layers can aggregate and funnel needs from layers above to layers below. In this model, as one goes up the stack, things become more customized.

Grand Challenge Communities

- One of the most appropriate means of developing a blueprint is to have the global scientific community agree on interfaces, and then to have the community compete on implementation
- Data and software, as well as research results, seem to be drivers of developing the blueprint
- But it is important to link those three with each other
- Publications of research results ought to be linked together with data sources and with software
- And all of these should be coordinated internationally, either by joint meetings of grand challenge science teams that would decide on global data policies and their implementations, and/or by Global Software Institutes and Centers of Excellence for each scientific discipline
- Need for “shepherds of global codes” to provide particular software services to global scientific community

Third Discussion: Suggested Actions (1)

- Big Science
  - Start a blueprint process
- Small Science
  - Create a presentation on small science that inspires an audience and creates a shared vision
    - Similar to Ed Seidel's presentation on big science; how it is changing, what is needed to enable it to continue to make scientific progress
  - Survey existing small scientists, focus on solutions that work, analyze why, look for commonalities
  - Longer term
    - Need simple-to-use workflow environment that supports easy definition of simple workflows
    - Need pervasive storage system and associated services

Third Discussion: Suggested Actions (2)

- Grand Challenge Communities
  - An international charrette-like process should be launched by a fairly small, but representative steering committee to define a blueprint for global CI
    - Consulting with all interested stakeholders is very important for the success of the process
    - EarthCube was mentioned as an example of a successful charrette process, although its scope was narrow – the US region only
      - The goal of EarthCube is to transform the conduct of research by supporting the development of community-guided CI to integrate data and information for knowledge management across the Geosciences
    - We need something similar for the global CI and all grand challenge community science fields

Outcomes

- Community appreciates the need to work together, encourage reuse, integration, cost-effectiveness
- Initial version of mission statement written, says process should
  - establish a globally coordinated framework for the development, integration, operation and maintenance of common infrastructure
  - establish a responsive requirements driven process that identifies the needed common infrastructure components
- Requires interaction between CI providers, service and tool providers, data services, network services, funding agencies, and users
- Interest from NSF and EC on synergies
  - EC in planning phase for Vision 2020 program
  - Conceptualizations for Software Institutes
  - Emerging data plans
  - EU-US coordination of grand challenge teams in FY14
  - International “charrettes” for community engagement and planning??? (Big data, long tail science)
  - Asian representatives following up with their communities

Proposal for a Blueprint Process (1)

- Should be open to all CI stakeholders (providers, operators, consumers, funders)
- Motivation
  - to provide sustainability to CI, and to ensure funding is well-spent and well-directed
  - to support new international grand challenge communities as well as ongoing science projects
  - to provide effective international cooperation
  - to provide a foundation for news aspects in data and software
- Will include elements such as
  - security assurances, identity management
  - data sharing policies
  - collaborative software support
  - campus bridging to international infrastructures
  - governance mechanisms
  - best practices
  - reuse
  - continued innovation

Proposal for a Blueprint Process (2)

- Need to balance the stability of the blueprint, as a well defined layered stack of interfaces, with the necessity to evolve CI along with changing science requirements and changing technologies
  - Thus there will not be one blueprint, but rather a blueprint process
- Should be driven (not dominated) by small, focused steering group
- Should follow a regular meeting cadence
- Support from funding agencies is essential, both practically (e.g., travel funding) and for process (e.g., running charrettes)
  - Funding agencies will need to implement parts of blueprint recommendations in order make it useful; failing that, the blueprint will remain an academic exercise
  - Funding agencies will have to support process by encouraging and supporting the participation of science communities
- Blueprint will represent an architecture template for CI
  - As blueprint will need to continuously evolve, the output of the process will be snapshots of that blueprint.
- Blueprint will likely be rendered as a layered stack of interfaces (vs. a stack of implementations and technologies)
  - Thus the blueprint could benefit from and drive standardization efforts
- Could use OGF infrastructure and process (mailing list, wiki, meeting cadence, etc) to implement blueprint process

Summary

- Workshop made a significant contribution toward establishing common vision among community leaders who have to do the work to unify the field, and among agencies who will fund the activities
- Still a long way to go, especially in international cooperation between funding agencies and scientific infrastructure providers
- Goal is developing common vision that is needed to support efforts:
  - Many fields have depended on computational science simulations, and many now are beginning to depend on computationally intensive data analysis
  - Infrastructure providers seek to build computational systems that support these researchers
- The meeting brought international experts representing such distributed scientific infrastructures as XSEDE, EGI, OSG, NAREGI, PRACE, FutureGrid, and PL-Grid, among others, but even broader participation in future activities is needed

Next Steps

- Need more application (user) community inputs
- The workshop was a followup to the SC 2011 BOF "Towards a Unified Cyberinfrastructure"
  - https://sites.google.com/site/toaunifiedci/sc11-bof
- Is part of a series
  - https://sites.google.com/site/toaunifiedci/
- Watch our activities and participate in the future workshops and ongoing discussion
- Currently seeking comments on these ideas, especially mission statement and proposed process
  - Draft workshop report circulated to workshop attendees, will be public in a few weeks
- All CI stakeholders are invited!
- Please send your comments to: grid-next-decade@lists.man.poznan.pl