B4 in E3SGS: Open Science Cyberinfrastructure Facilities, and Distributed Systems

Doug Olson (dlolson@lbl.gov)
Lawrence Berkeley Laboratory

Viewpoint

- B4. Develop tools and methods to protect the distributed information technology infrastructure: ensuring network security, preventing disruption of our communications infrastructure, and defending distributed systems against attacks.
- Breakout session goals:
  - Definition of state-of-the-art and futures
  - Identify gaps or bottlenecks (to achieve?)
  - Identification of research programs to address the gaps
- **Consider the open science**
  - large cyberinfrastructure facilities with large numbers of users.
  - distributed systems and the networked computing environment.
- Also
  - Exascale implies sharing resources between open science and users with proprietary data needs

Context

- The applications will be large collaborative efforts at a scale nearly unprecedented for computational science
  - Integrating large complex models and managing the produced information requires tightly coupled teams of researchers each with expertise in various aspects of the models and data.
    - ✓ Many computational scientists are not used to this and some sociological changes are required.
    - ✓ Some additional cyberinfrastructure functionality is implied.
  - Many of the collaborations will have international scope.
    - ✓ International open science collaborations with large membership often have ***important*** and ***different*** identity management and privacy requirements than the parochial view of a single DOE lab.
    - ✓ Implies adjustments to policy and related cybersecurity infrastructure functionality.
  - Remember Dave Bader's comments on the need for functionality to support collaboration (after all, the goal is SCIENCE, not just flops)

Functional Areas

- Identity Management (IdM)
- Authorization/Privilege Management (AuthZ)
- Collaborative Tools (Collab)
- Application (user) monitoring (UserM)
- System monitoring (SysM)
- Network monitoring (NetM)
- Usability monitoring (UseM)
  - a different axis – a requirement on solutions
  - Solutions should not just prevent bad behavior, they should enable good behavior

State of the present & future art

- IdM, AuthZ (identity and privilege management)
  - In HPC – validation & unix uid/gid at facility scope
  - In Distributed Computing – X509, VO management, mapping grid-wide IdM attributes to facility features
  - Future
    - ✓ current grid IdM functionality ***goals*** serve needs for international collaborations at exascale
    - ✓ Infrastructure & policy frameworks still need work – to achieve known goals as well as ensure performance needs at exascale
- Collab (collaborative tools)
  - Wiki, listserv, sourceforge, video conf, Skype
  - Future
    - ✓ Couple to IdM, AuthZ
- UserM (are users' behaving suitably)
  - Some level of file system scanning and heuristics on authentication
    - ✓ in other words, almost nothing
  - Future
    - ✓ Executing an application using large resources in a sizable collaboration needs some QA and validation by the collaboration to authorize the usage
    - ✓ Monitoring for bad behavior
    - ✓ In large collaborations the range of improper behavior is not very different from the general public
    - ✓ Virtualization? (maybe not on exascale machines)

State of the present & future art

- SysM (system monitoring)
  - File system scans, authentication heuristics, auditing, log analyses
  - Future
    - Exascale systems - ???
    - Distributed Systems – distributed auditing, logging and analysis
- NetM (network monitoring and control)
  - Intrusion detection (BRO, commercial implementations)
  - Firewalls
  - Future
    - More complete deployment of BRO-like passive monitoring with active feedback
- Use (usability monitoring of implementations)
  - Huh?
  - Future
    - Have a coherent model across DOE open science facilities and testbed to validate implementations before deployment?

Gaps & Bottlenecks

- IdM
  - Missing coherent model, not just implementation, across DOE open science resources
  - I2/Shibboleth has model for universities, needs some work to see how it could work for DOE facilities
- AuthZ
  - Missing coherent model across DOE facilities
  - Grid domain efforts show promise
- Collab
  - Only limited infrastructure is generally available
- UserM
  - Needs requirements gathering on how to determine acceptable behavior, as well as improper behavior
  - Infrastructure supporting more secure user written software
- SysM
  - Must be some
- NetM
  - Must be some (BRO everywhere?)
- Use
  - Needs testbed as well as workflow descriptions to allow testing

Research Topics

- IdM
  - Develop coherent model for DOE open science facilities & international collaborations – *Is GridShib enough?*
  - Investigate performance issues for exascale – *will data and/or cpu parallelism eventually cause authentication bottlenecks?*
- AuthZ
  - Develop coherent model for DOE open science facilities & international collaborations – *Is grid (OSG/EGEE) infrastructure (VOMS, GUMS) coupled to GridSHib enough?*
  - Investigate performance issues for exascale – *will data and/or cpu parallelism eventually cause authorization bottlenecks?*
- Collab
  - Integrate open source tools with IdM, AuthZ
  - Deploy functional infrastructure
    - ✓ is it bureaucratic research (just gimme $), or CS research?
- UserM
  - Couple to AuthZ
  - Digitally signed applications?
  - Distributed logging, auditing and analyses
  - Virtualization

Research Topics

- SysM
  - Distributed auditing, logging and analyses
  - Virtualization
  - Statistical Learning Theory and Control Theory
- NetM
  - BRO-like monitoring and cleaning at the ISP level rather than facility level?
    - ✓ Think of ISP like water company. It provides clean water to all houses and each house only does a little additional filtering where desired. Each house does not have a full water processing plant.
  - Monitoring at the exascale data rates?
  - Statistical Learning Theory and Control Theory
- Use
  - Integration testing and deployment strategy for DOE open science networked computing environment