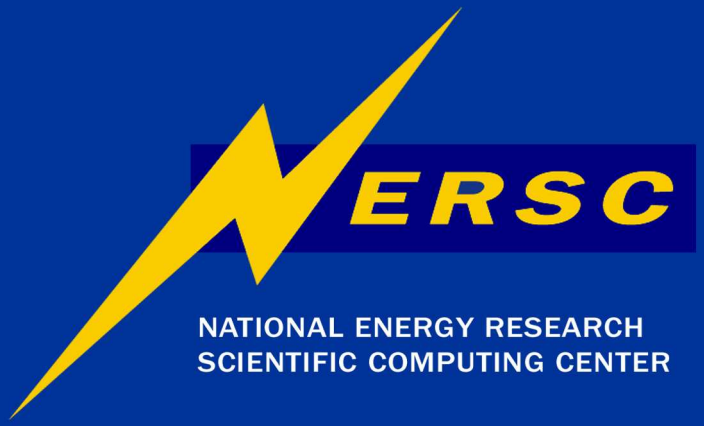


Publication and Protection of Sensitive Site Information in Grids

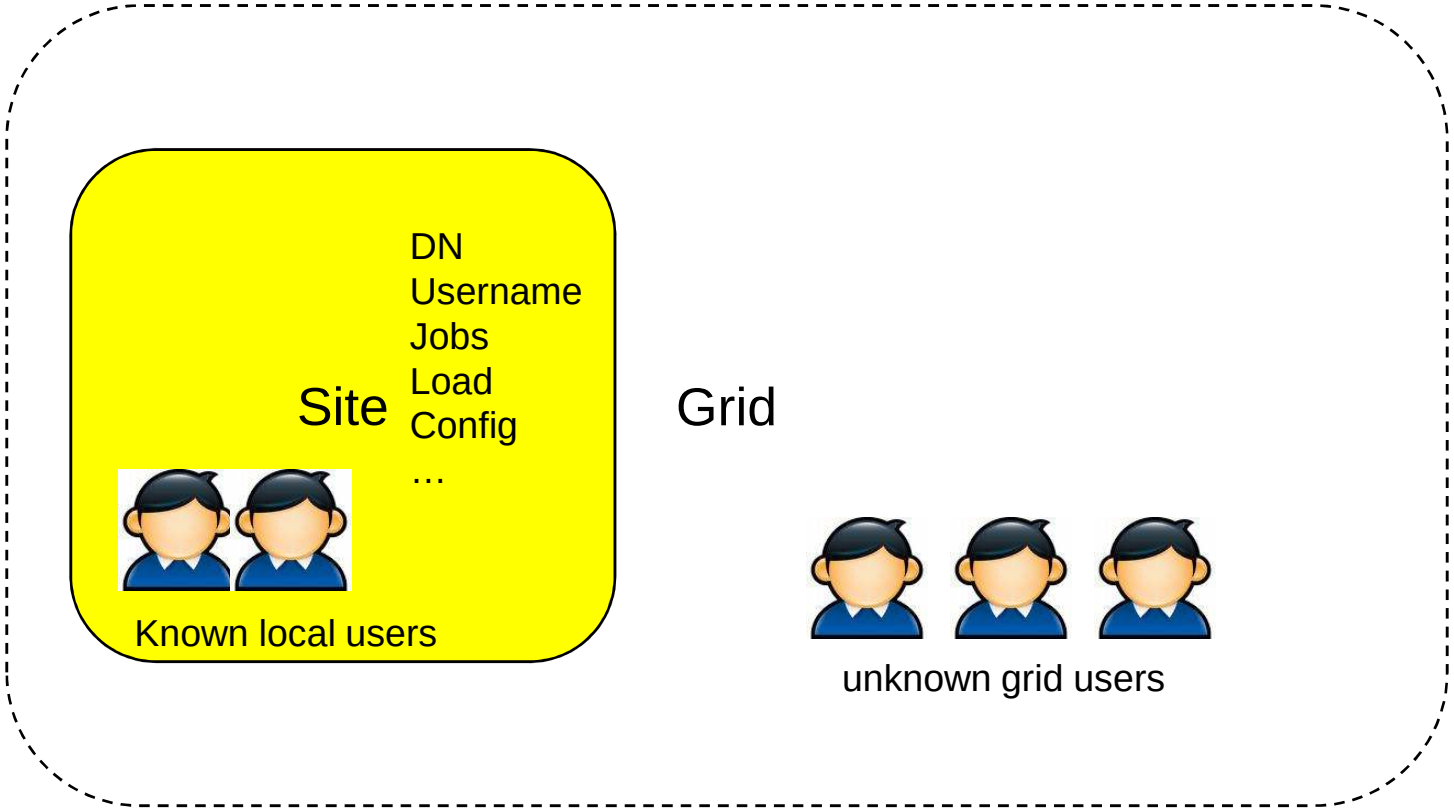
Shreyas Cholia <scholia@lbl.gov>
NERSC Division, Lawrence Berkeley Lab

STPG Workshop - CCGrid 2008
May 22nd, 2008

Information Collection in Grids

- To create a successful and functional grid you need to collect information from sites
- Grid infrastucture must publish collected information and make it available to interested parties
- We want to analyze the vectors of information collection
  - Systems publishing/collecting information
  - Type of information being gathered
  - Methods of data protection applied to this information

Information Boundaries



Focus on Open Science Grid

- What is the Open Science Grid?
  - Virtual Facility providing distributed compute and storage resources
  - Comprised of VOs and their Users + Resource Providing Sites + OSG Infrastructure Providers
  - Broad range of sites - small universities to large national labs
  - Must have flexible infrastructure to meet diverse site/VO requirements
  - Information collection and publishing coordinated by Grid Operations Center (GOC)
- NERSC/LBL heavily involved in OSG
- Our study started out as a recommendation report for OSG, but many of the results applicable to other grids

Open Science Grid

Office of Science
U.S. DEPARTMENT OF ENERGY

Information Being Published

- Resource Selection Information
- Monitoring
- Accounting
- Troubleshooting
- Log Files
- Site Availability information
- Site Validation

Information Collection Systems in OSG

- GIP/CEMon
- Gratia
- Syslog-NG
- RSV
- site_verify
- Monalisa
- Others?

CEMon

- Periodically queries Compute Element state
- Publishes CE information as GIP attributes
- Information made public through BDII and ReSS Condor Class-Ads
- Used for resource selection queries

CEMon Sensitive Info

- Operating System version info
- Underlying jobmanager
- Internal System Paths
- Authentication Method

All this is necessary for a successful grid query BUT:

- Site must understand that info is public
- May want to restrict level of detail to avoid a "Google hack"

Gratia

- OSG Accounting System
- Sites install local probes that report job/storage usage records to collector
- Information published through web interface
- Web interface supports custom SQL queries

Gratia - Sensitive Info

- User DN and local account names
- Job Information

Risks:

- Users may consider job information private.
- If DN (or password) is compromised, it becomes very easy to discover other sites supporting the same DN.

Syslog-NG

- Collects grid log files at a central collector
- Centralized Log Collection
  - Troubleshooting distributed grid workflows
  - Security Incident Response
    eg. where was a compromised DN used
- Queriable database backend
- Tiered architecture

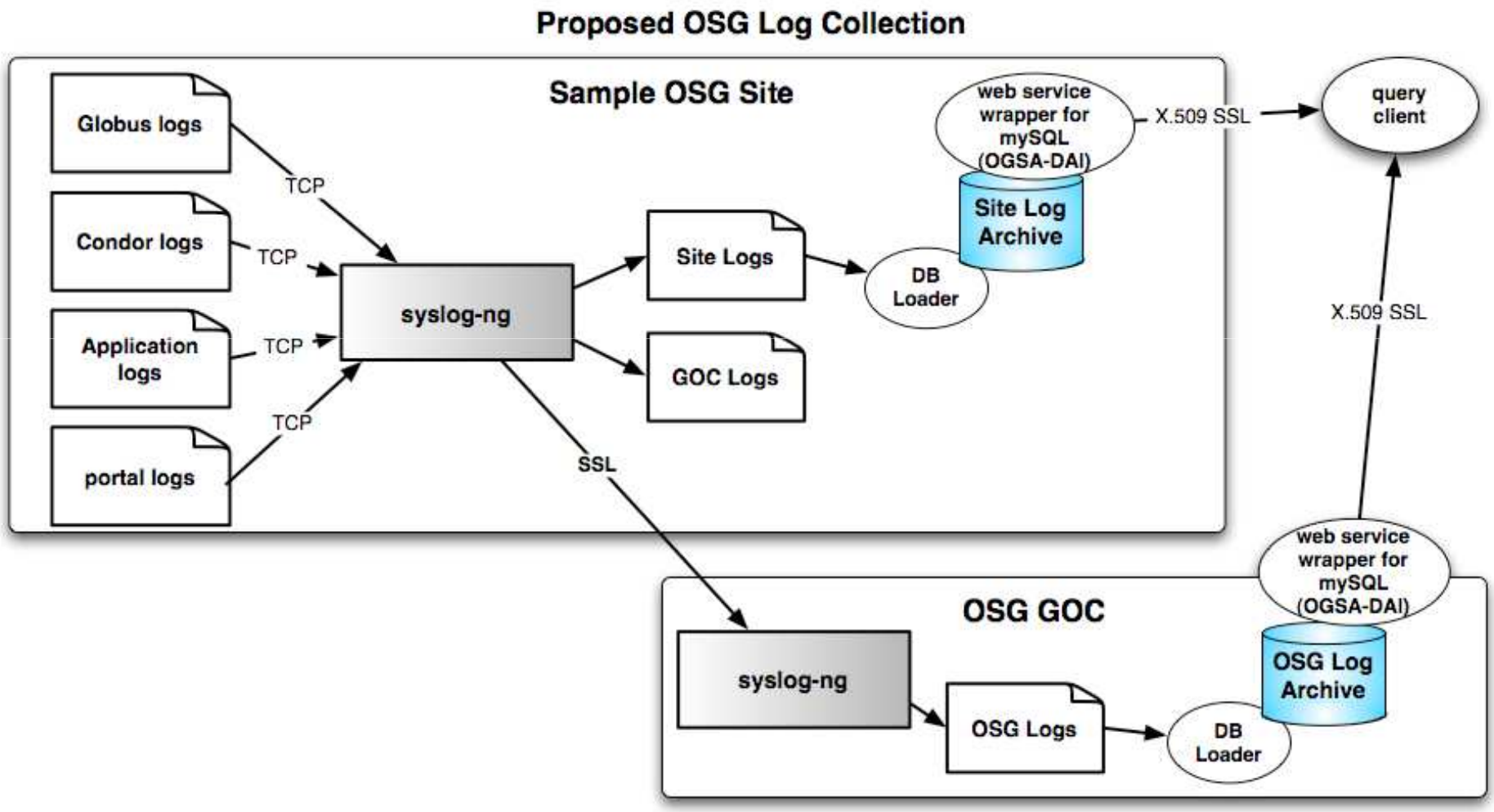
Syslog-NG Risks

- Log files are sensitive! Most sites want to limit access to these.
  - Internal system info - may expose vulnerabilities
  - Detailed user, software info and failure modes
- May not want to make these available to grid infrastructure providers
  - No longer under site control

Answer:

- Tiered architecture design allows sites to set up local collectors that can filter and forward limited information to the grid

Tiered Logging Solution



Monalisa

- Publishes
  - resource availability
  - load information
  - Performance information
- Public web interface. Very useful for querying the "state of the grid" at a high level: http://monalisa.caltech.edu/monalisa__Repositories.htm

BUT

- May be used to target overloaded sites for DoS

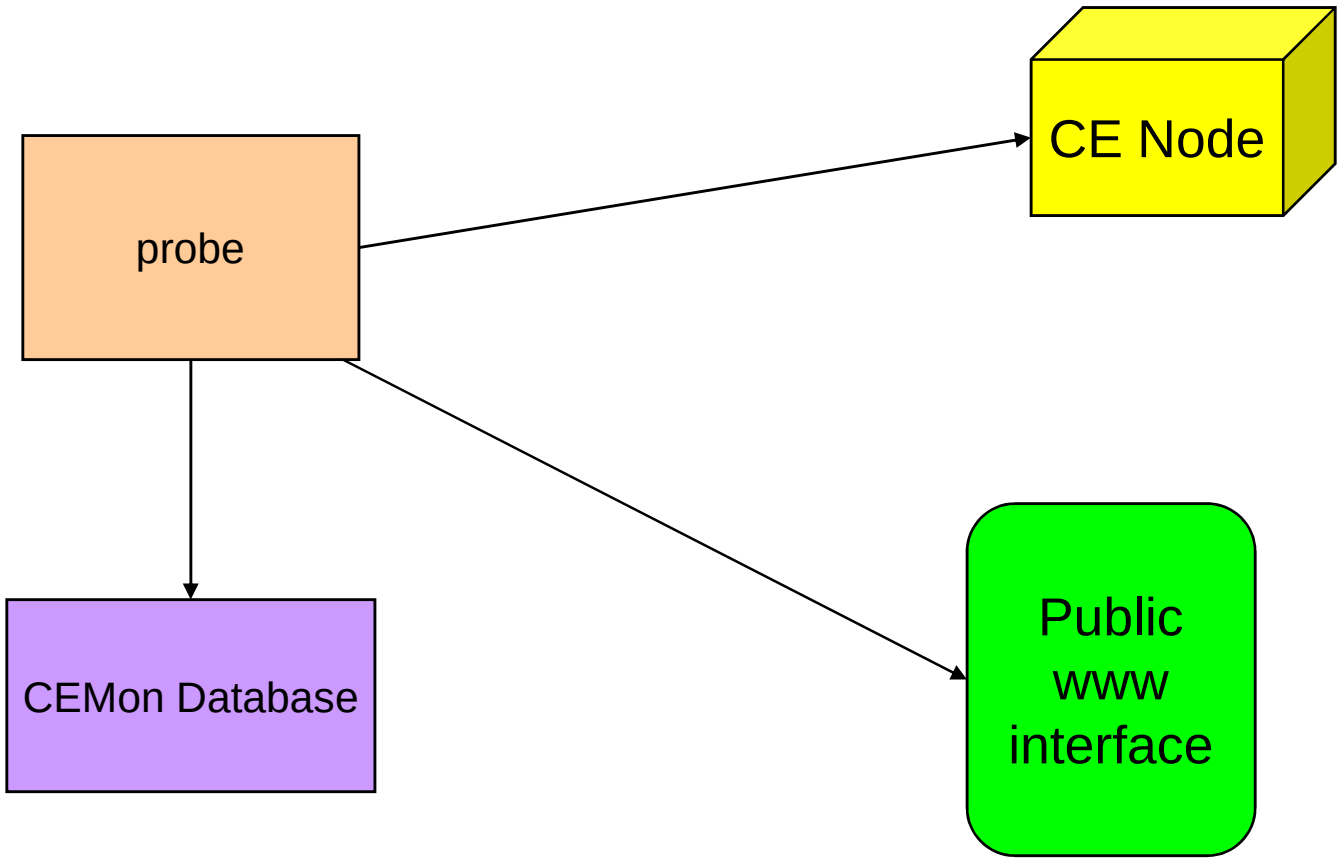
RSV and Site-Verify

- Probes run by grid infrastructure to verify site capabilities and report on site availability
  - Verifies information published by CEMon
  - Publishes results online: http://vors.grid.iu.edu/cgi-bin/index.cgi

Risks:

- Same risks as CEMon info
- Additionally, historical data is available - may be able to trace downtimes (when system is in transitional state).

RSV/Site Verify Publishing



Summary of Sensitive Info

- Account Names, User DNs (VORS, Gratia)
- Failure Modes, Security Related Details (Syslog-NG)
- Historical System Availability (VORS)
- System Load (MonALISA)
- Application Names, Internal Paths (Gratia, CEMon)
- Software Levels (CEMon)

Security Risks

- Gratia - public interface to Gratia DB
  - Track user activity on a site
  - Rival project can discover job information
  - In case of compromised cert/account, query DB for other sites with same account
- Syslog-NG
  - Internal failure modes, other logging details available to non-site personnel
  - Security incident details no longer private
- CEMon/VORS
  - List of valid user accounts, DNs made public
  - Software levels, Authn method public - possible “Google Hack”
  - Historical archive of system info (may be able to target recurring downtimes)
- MonALISA
  - System Load Info - DoS attack during high load

Data Protection (for Sites)

- Turn down logging in Syslog-NG to minimal level
  - Start-stop times, User DN info
  - Increase level for troubleshooting
- Customize probes to meet site requirements.
  - Only publish necessary information
  - Be AWARE of what is going out!!
- Modify GIP attributes
  - Override or Modify attributes as necessary
- Mask sensitive data
  - Use generic VO names instead of local account names
- Site level collectors
  - Review and filter, before forwarding to OSG
- Choose secure/encrypted publication channels

Suggestions for OSG

- Authenticated access to information services
  - Use GSI certs within browser to authenticate user
  - Limit access based on VO
- Consolidate services where possible
  - Minimize information streams publishing the same data
  - Teragrid INCA as a model?
- Use encrypted SSL based communication for ALL information streams
  - https, GSI etc.
- Use “robots.txt” to prevent web caching
- Authenticate probes using GSI hostcerts to prevent bogus information.

Conclusions

- Not a replacement for hard security policies
  - Must fix and patch software regularly
  - Internally monitor systems
- Sites should have more flexibility and control over published information
- OSG should consider limiting public access to user/VO based access