Illuminating the Shadow Mesh:

Red Storm Visualization

Warren Hunt

Sandia National Laboratories

Albuquerque, New Mexico

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Illuminating the Shadow Mesh

Red Storm Overview

- 5 cabinet rows, 175 total cabinets
- 12,960 dual-core compute nodes: 25,920 total cores
- 640 dual-core service nodes
- 101.4 TFlops performance, #2 on Nov. ’06 Top500

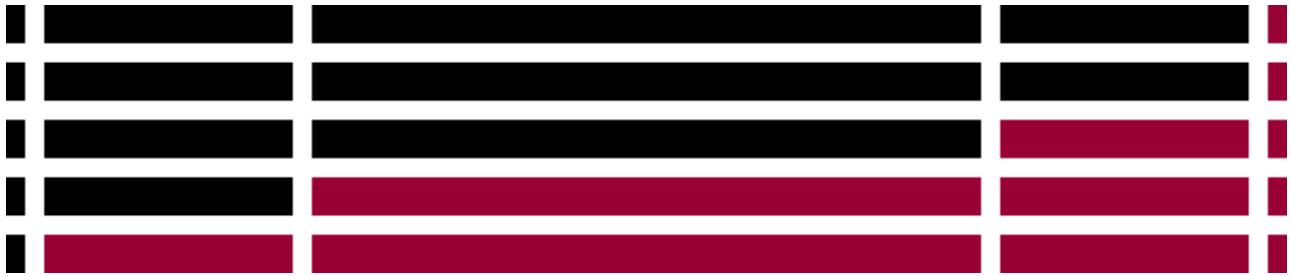
LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Red Storm Overview

- 320 TB of disk storage
- Lustre file system
- Partial Torus
- Lightweight Catamount kernel
- Physically re-partitionable



LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz Design Considerations

- Represent the logical or virtual system
- Represent the physical system

- Mapping Information, Wish to overlay:
  - Physical onto logical
  - Logical onto physical
  - System meta-data onto visualization
- Employ 3D visualization to explore & understand routing

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz Design, User Profiles

- System Administrators & Managers
  - Need status & statistics
  - Need real-time decision support
- System Engineers
  - Need detailed, low-level information
  - Need physical-to-logical system translation
- XT3 Developers
  - Need statistics and historical information
- HPC Research & Analysis
- General HPC Management

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz, Goals for base functionality

- Must present multiple views of system
  - Physical View
  - Logical View
- Tight coupling with system logs and data bases
- Provide interface for direct MySQL querying
- Luz user interface (GUI) must be intuitive & powerful
- Leverage information visualization techniques
  - Present mesh as a volume
  - Transparency for context without clutter
  - Interactive object selection

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz, Base Functionality Support

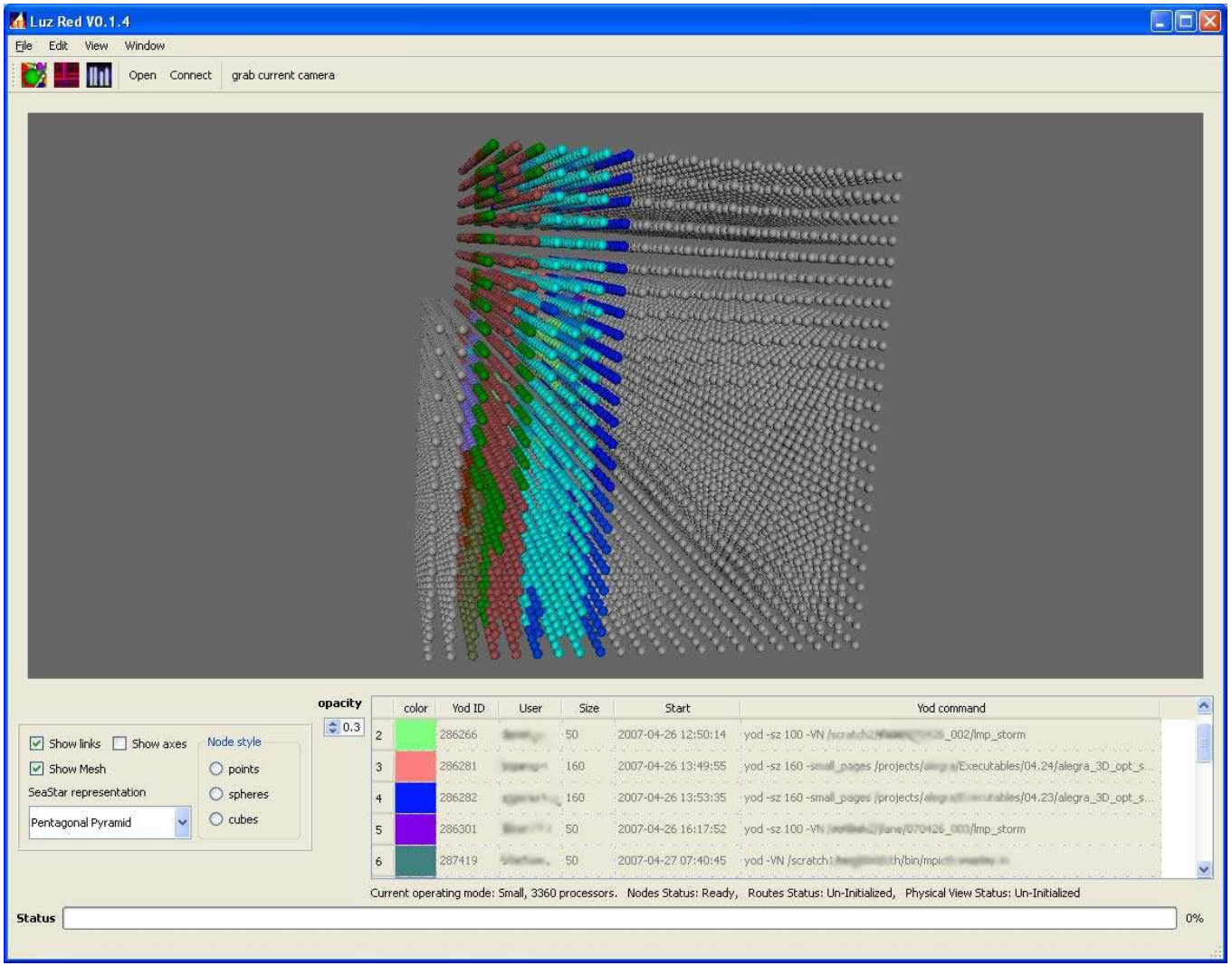
- Many layers below the hood
  - C++ Implementation
  - Visualization Tool Kit for visualization library
  - Qt for cross-platform GUI

- Luz is a suite of tools wrapped as one application
  - Scripts and daemons interacting with system data base
  - Network and firewall negotiation
  - Daemons monitoring system log
  - Job status and allocation updating
  - Scripts interacting with Luz database for ShadowMesh
  - Other Luz support services, address translation etc…

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz Preview

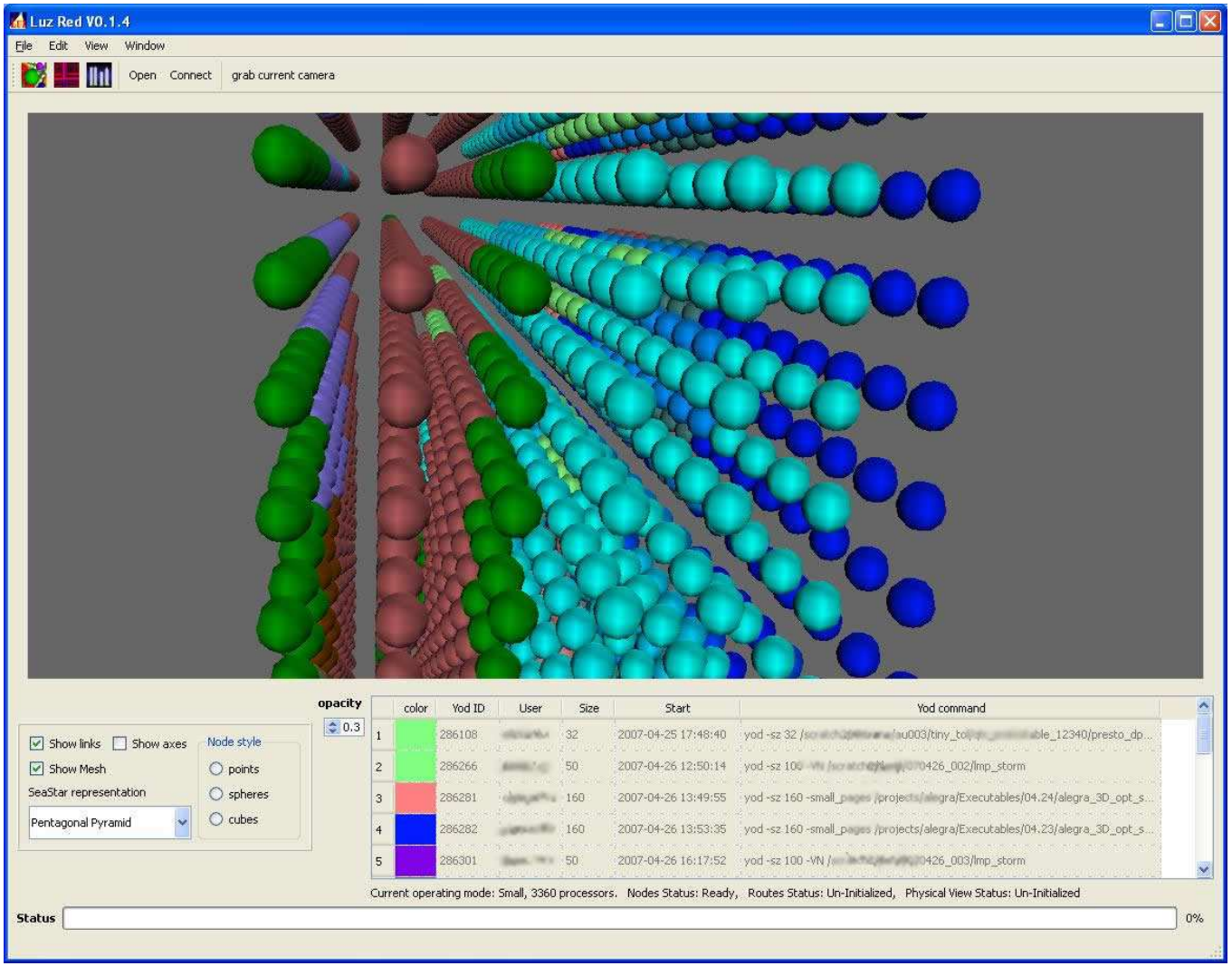


LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz Preview

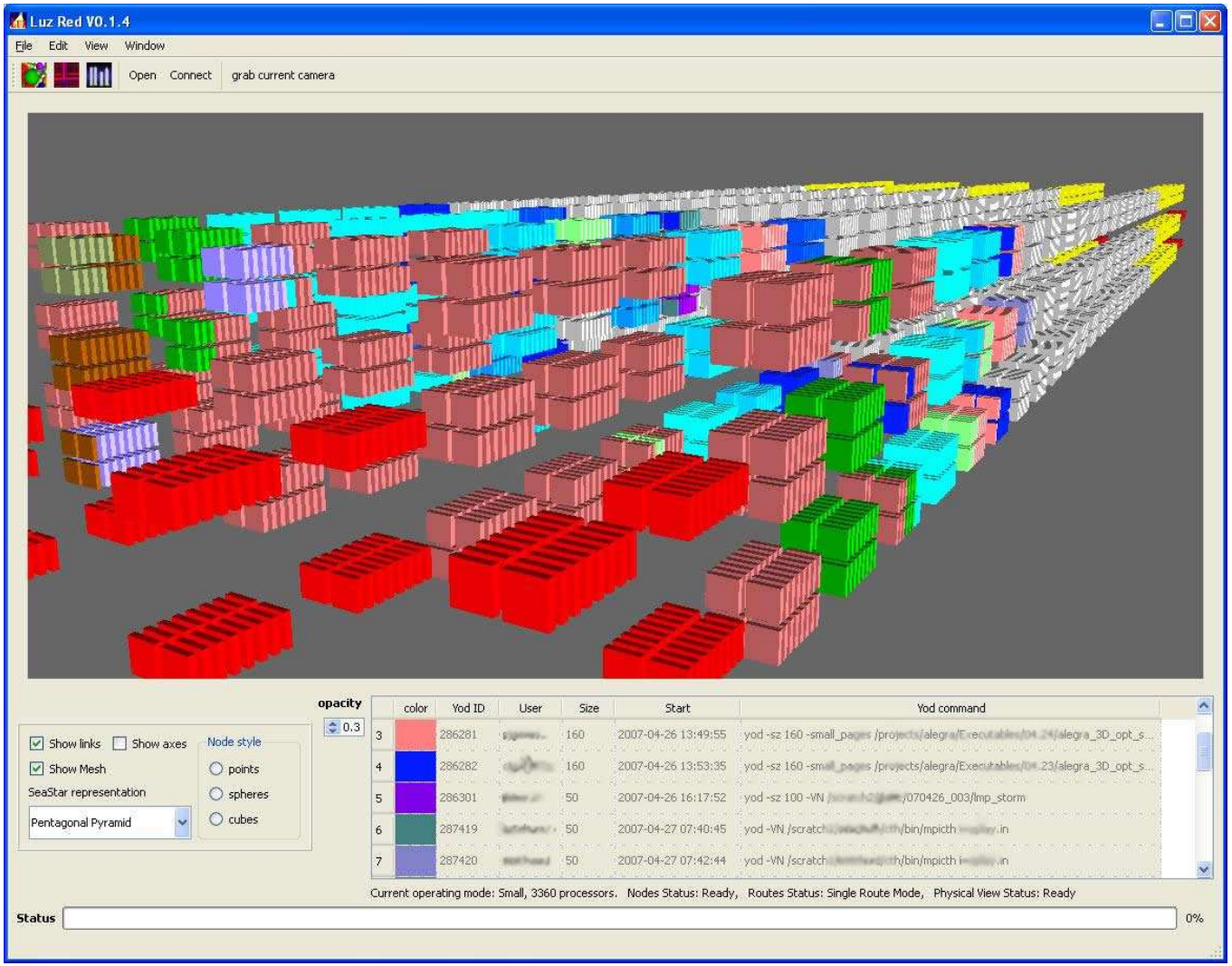


LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz Preview

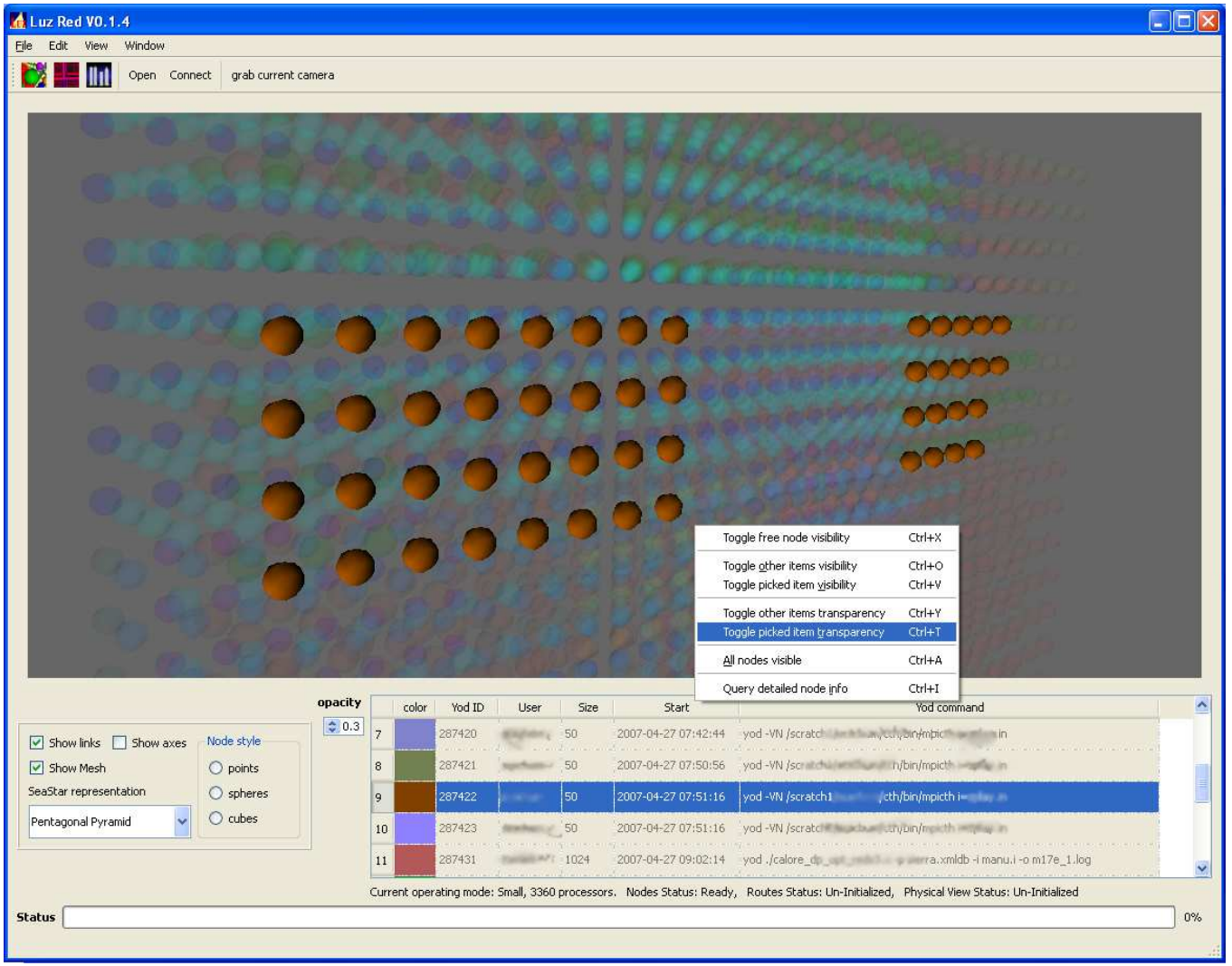


LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz Preview



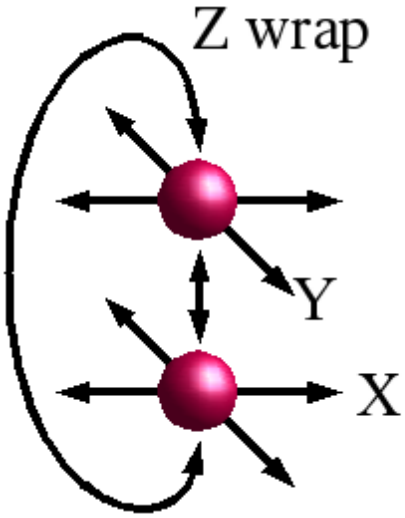
LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Red Storm Routing

- Logical space is 3D partial torus
- Cray SeaStar Networking
  - Bi-directional in (X, Y, Z)
  - Provides link to L0 control system
  - Also DMA engine
- Only logical Z dimension wraps
- Routing is static, calculated at boot
- Route dimension preference: X, Y, Z
- Big source of job failures, job hangs, maintenance concerns, unknown system state



LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz, Looking into the Mesh

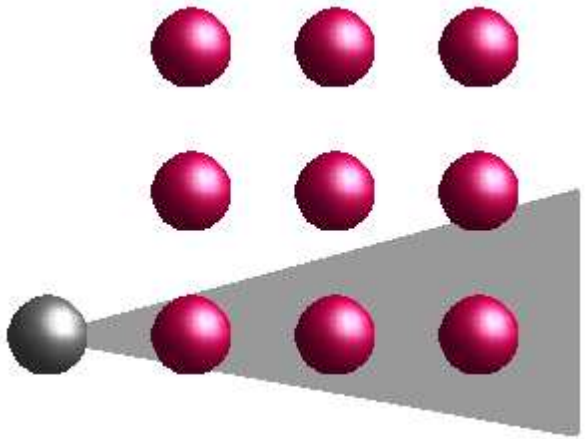
- Opteron failure
  - Current job needs re-start
  - Removed from node queue
- SeaStar failure
  - Current job needs re-start
  - What else? Good question!
- The failed job is explicit
- What are the implicit and indirect consequences?

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz, Shadow Mesh

- A dead SeaStar casts a ‘quiet’ shadow
  - No communication to PBS nodes
  - No communication to Lustre
  - No nuthin’
- Not able to dynamically re-route around the dead link
- All nodes downstream are now isolated and “in shadow”
- Question: downstream with respect to which nodes?
- What to do? Just re-boot 25K nodes?



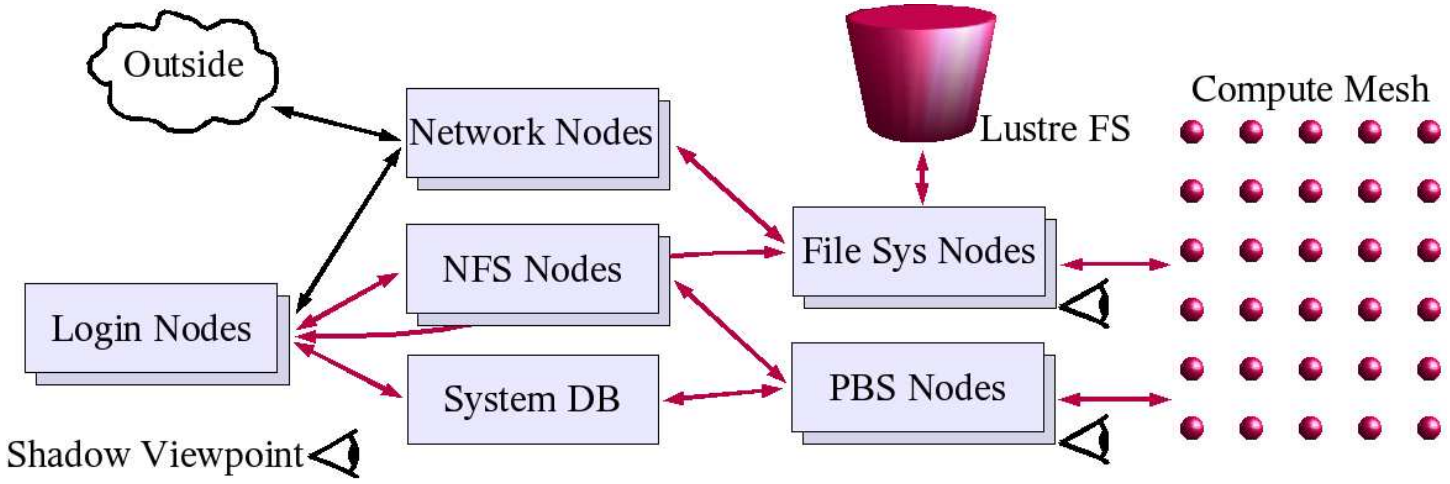
LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz, Defining a Shadow

- Many shadows cast in Red Storm by a dead link
- Potentially compute a “shadow set” from
  - Login nodes
  - PBS nodes
  - Lustre nodes
  - Compute nodes



LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz, Defining a Shadow

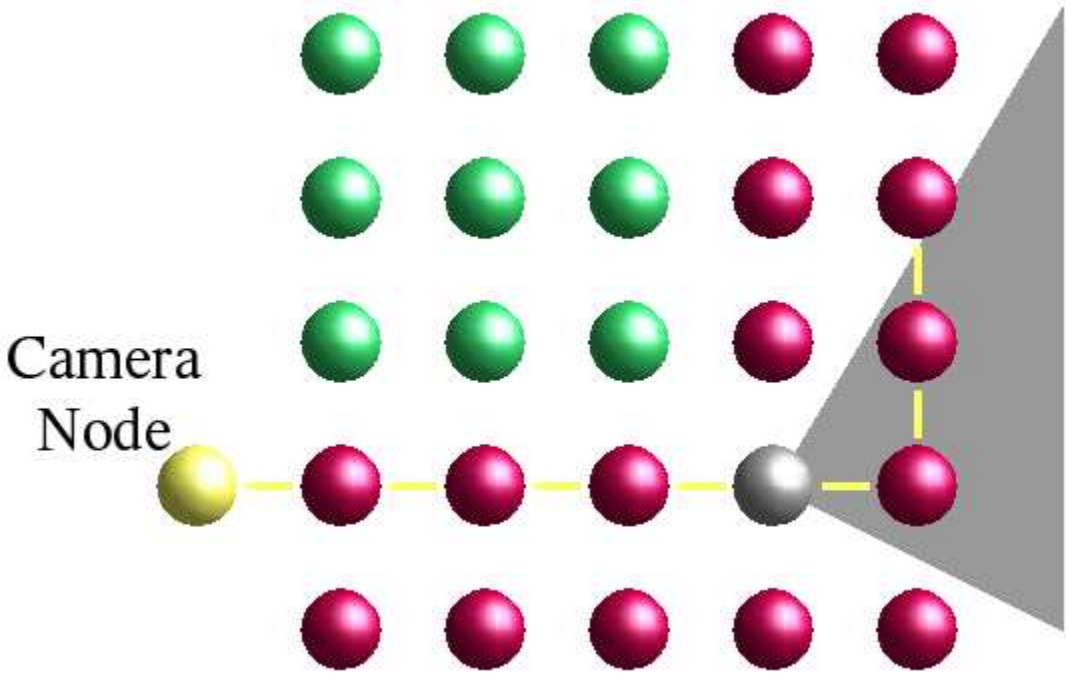
- Could define shadow as set of nodes suffering any reduced connectivity as result of dead link
- Motivation:
  - Enable decision making
  - Optimize Red Storm utilization and performance
- Shadow Mesh gives a cost

- System administrators use cost-benefit analysis to determine response to SeaStar failure
- Q: What is the shadow? A: That depends.
- Allow selection of “camera viewpoints” to define shadow

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz, Defining a Shadow

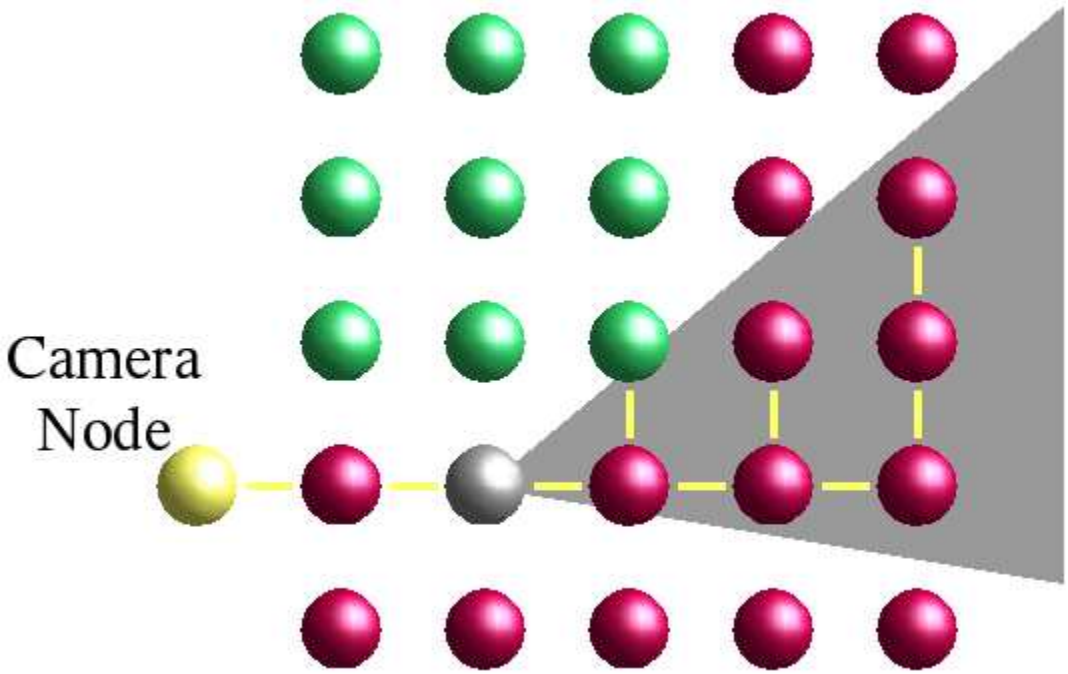


LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz, Defining a Shadow



LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz, Computing the Shadow Set

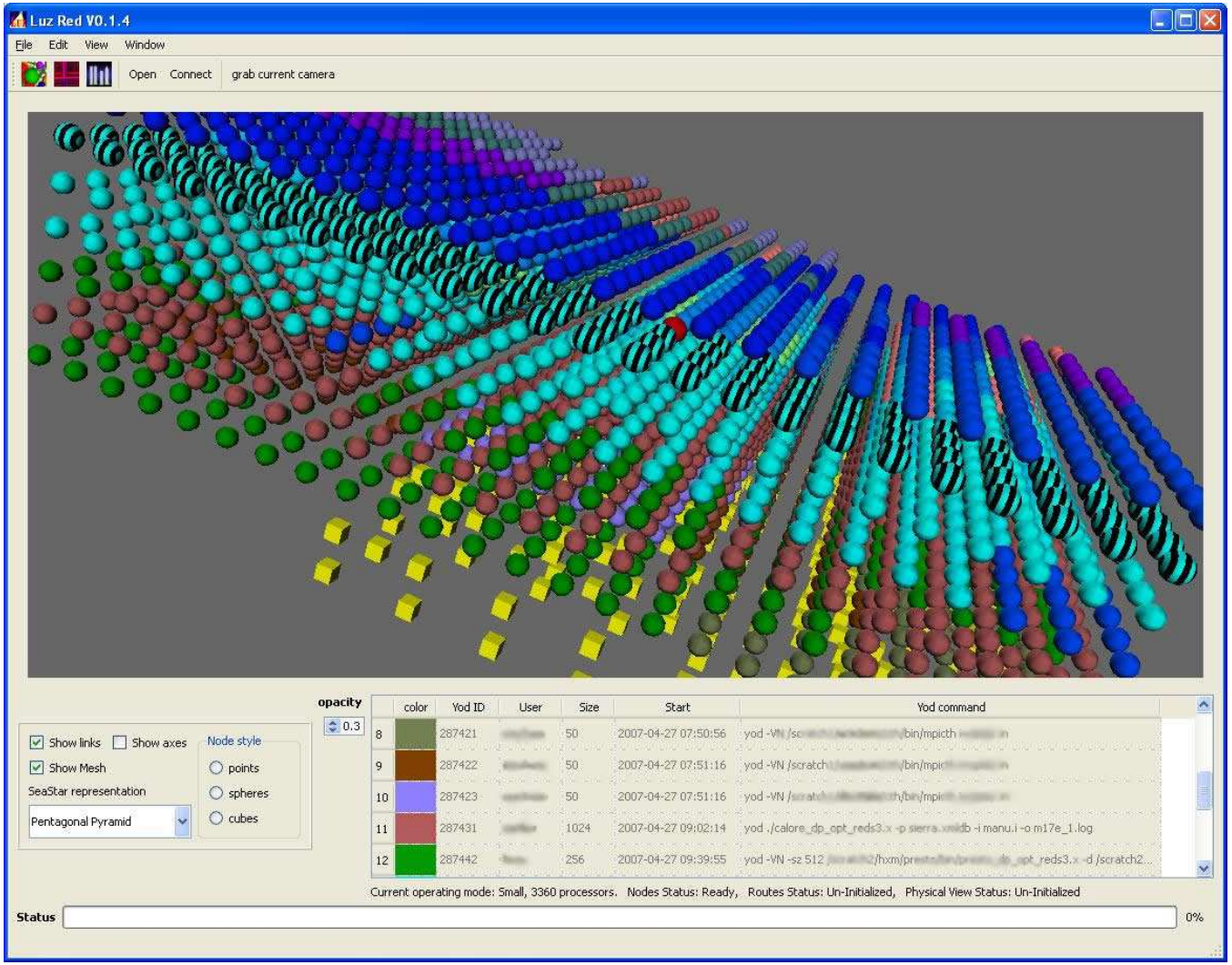
- Need custom MySQL schema to allow real-time shadow computation
- Routing is not symmetric
- Route table re-computed upon boot

```
+---------------------+-------------+------+-----+-
| Field               | Type        | Null | Key |
+---------------------+-------------+------+-----+-
| beg                 | smallint(6) | YES  | MUL |
| end                 | smallint(6) | YES  |     |
| hop_src             | smallint(6) | YES  | MUL |
| hop_dst             | smallint(6) | YES  | MUL |
| hop_src_output_port | tinyint(4)  | YES  |     |
| hop_dst_input_port  | tinyint(4)  | YES  |     |
+---------------------+-------------+------+-----+-
```

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz, Shadow Sample

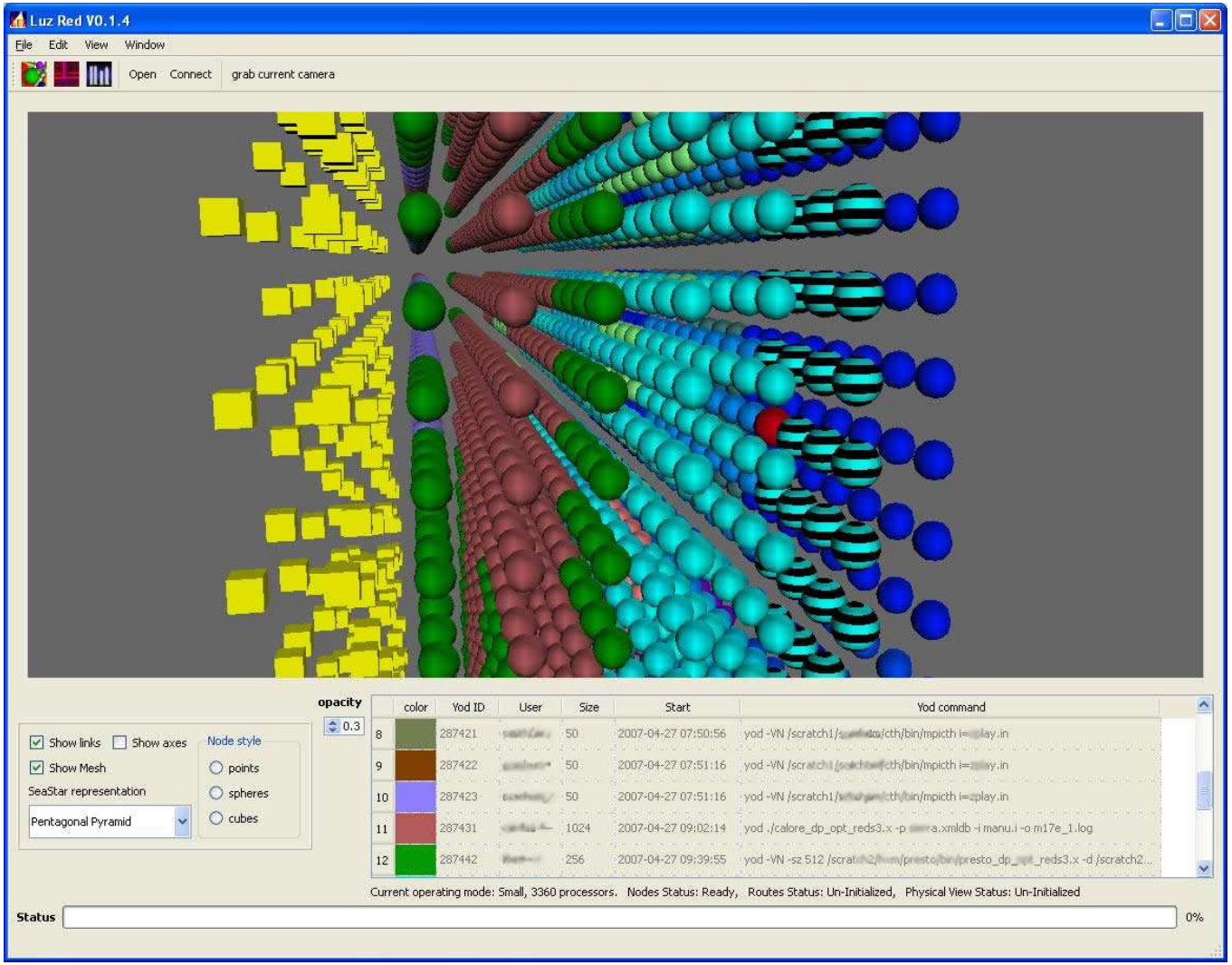


LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz, Shadow Sample

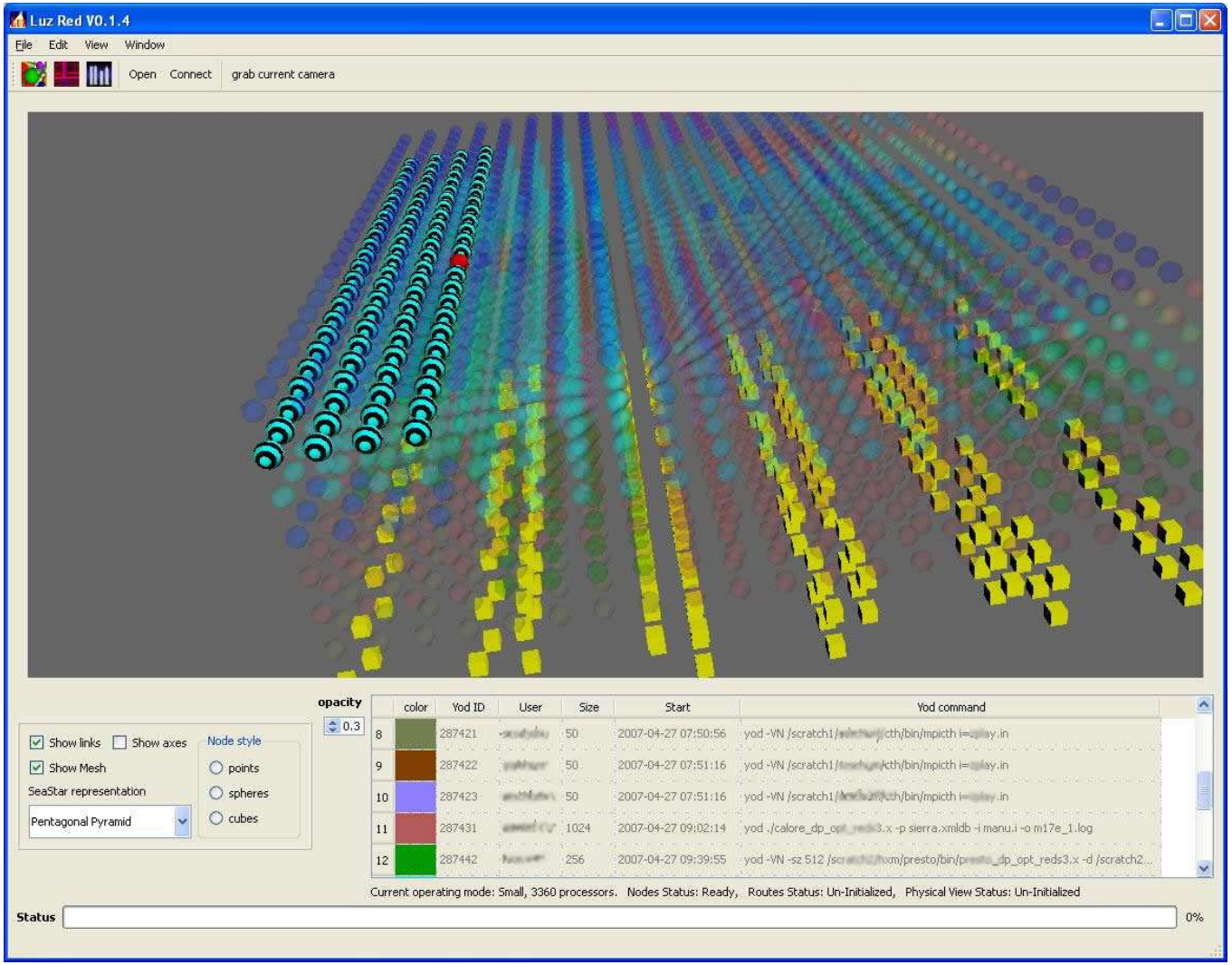


LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz, Shadow Sample

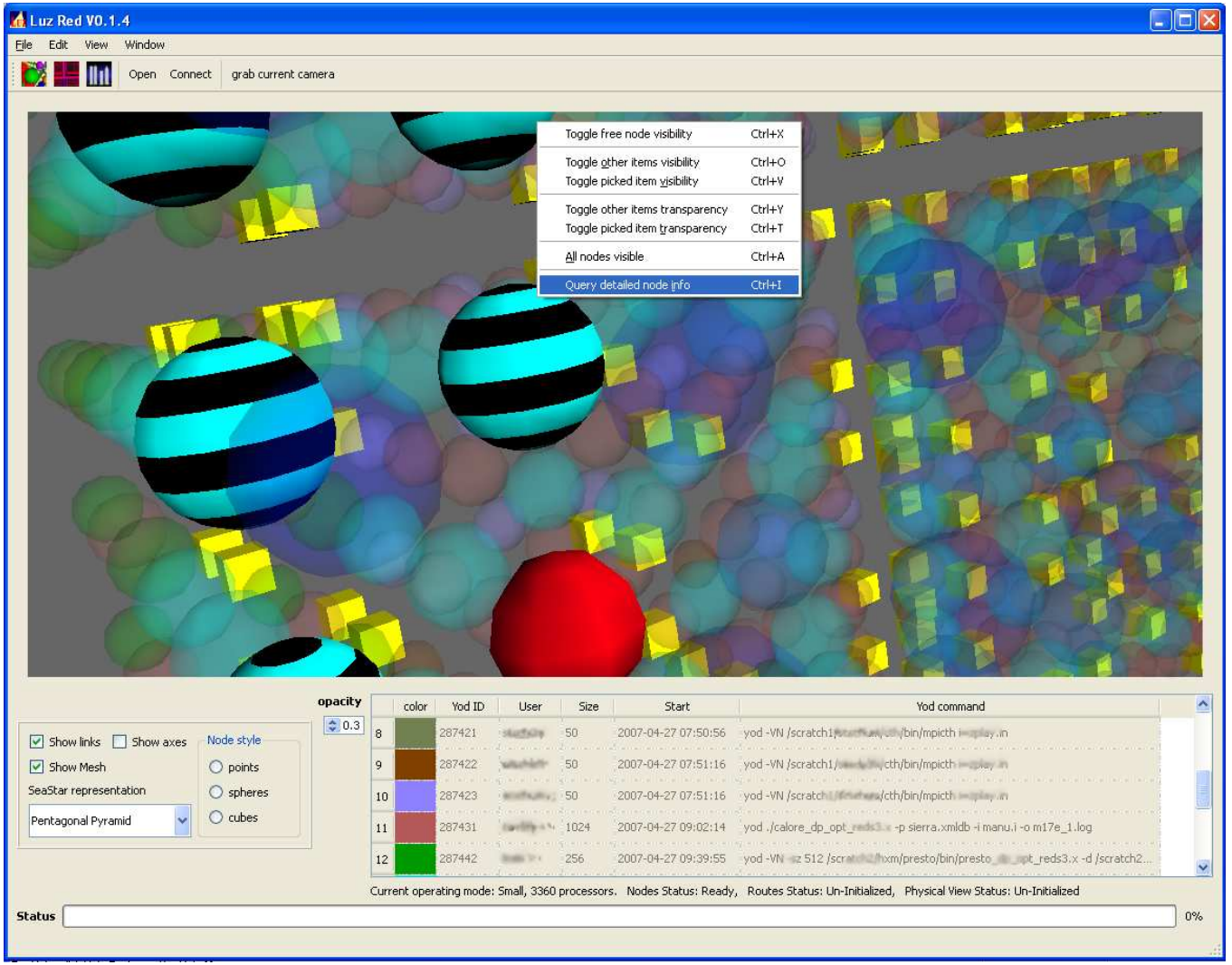


LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz, Shadow Sample



LUZ

Sandia National Laboratories

Illuminating the Shadow Mesh

Luz Conclusion

- Future
  - Shadow Islands & Shadow Refinement
  - Better torroidal representation
  - Improved annotation and ‘pop-up’ meta-data

- Thank you

LUZ

Sandia National Laboratories