Wide Area Cluster Monitoring with Ganglia

Federico Sacerdoti, Mason Katz, Matt Massie, David Culler

IEEE Clusters 2003 Conference

Hong Kong

Ganglia Cluster Toolkit
http://ganglia.sourceforge.net

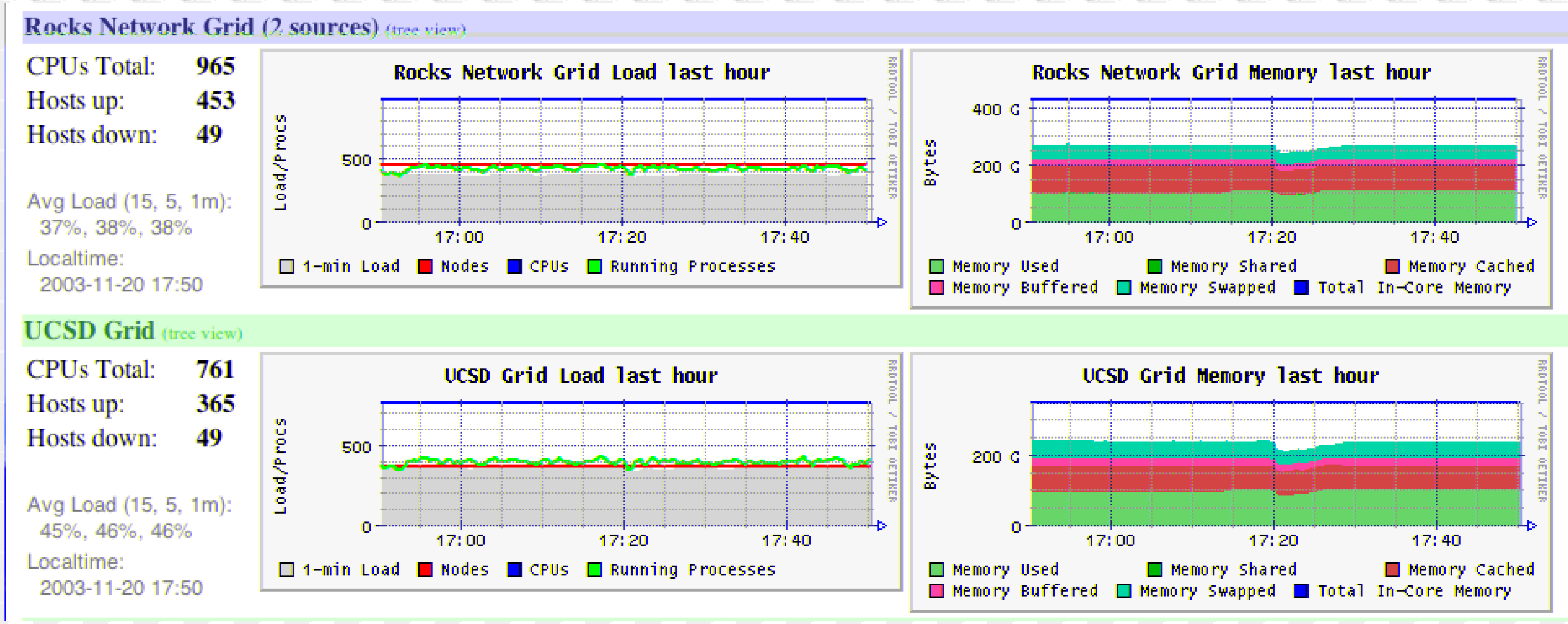
Introduction

- Monitoring Clusters with Ganglia
  - alive heartbeat,
  - cpu_load, mem_free,
  - bytes_in, bytes_out
- 1000-node cluster: **done**
- 10,000 nodes over groups of clusters: **our problem**

Ganglia Cluster Toolkit
http://ganglia.sourceforge.net

Contribution

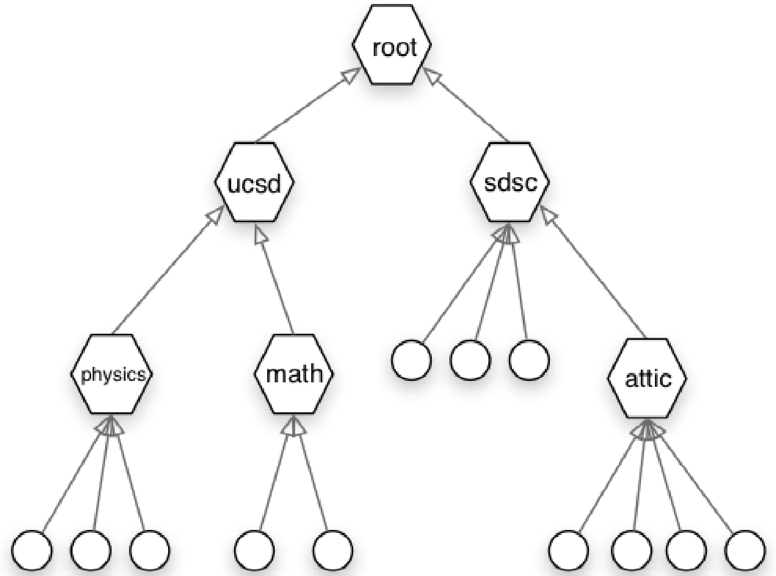
- With our technique, you can monitor 1,000,000 nodes, not be overwhelmed, and still have meaningful data to show.



Ganglia Cluster Toolkit http://ganglia.sourceforge.net

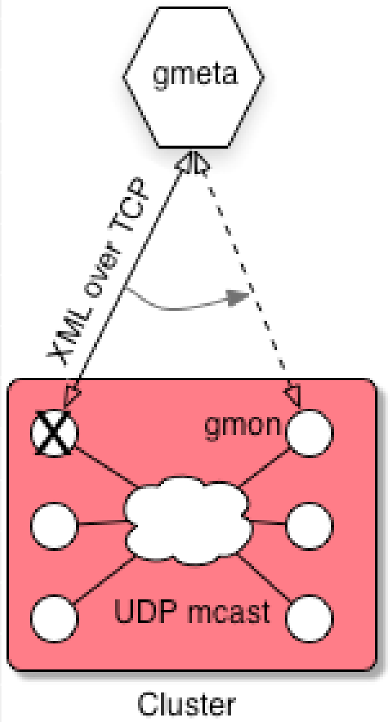
Contribution

- Monitored nodes can reside in one large cluster or many grid endpoints.
- Monitoring load is split between N ganglia agents.
- Introduce notion of a *Monitoring Tree* to handle load.



Ganglia Intro

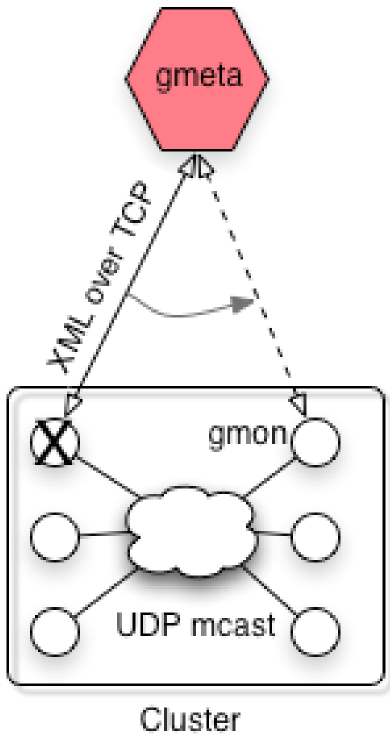
- Ganglia has two components: *Gmond* (within cluster), *Gmetad* (between clusters).
- **Gmond** runs on each cluster node.
  - Lightweight
  - UDP multicast to publish “metrics”
  - Each gmond agent has global cluster knowledge.
  - Isomorphic output from all nodes allows for easy failover.



Ganglia Cluster Toolkit http://ganglia.sourceforge.net

Ganglia Intro: Gmetad

- **Gmetad** runs on a management node.
  - Relatively Heavyweight.
  - Collects monitoring info from gmond.
  - Keeps metric histories over time.
  - Uses TCP to pull XML Ganglia data over the wide area.

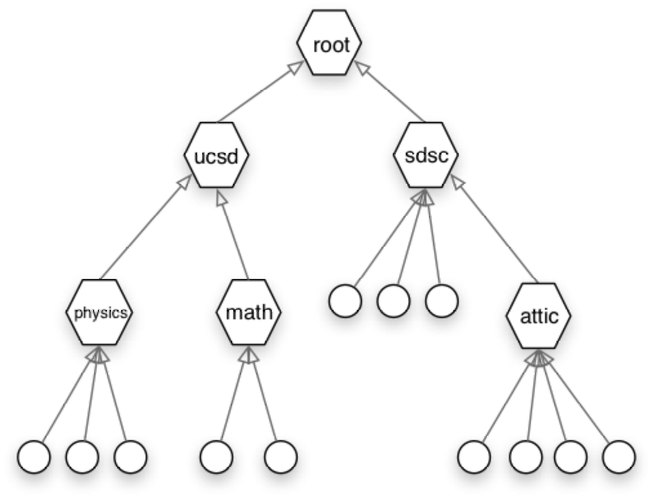


Ganglia Metrics

- Metrics are monitored by gmond, presented by gmetad.

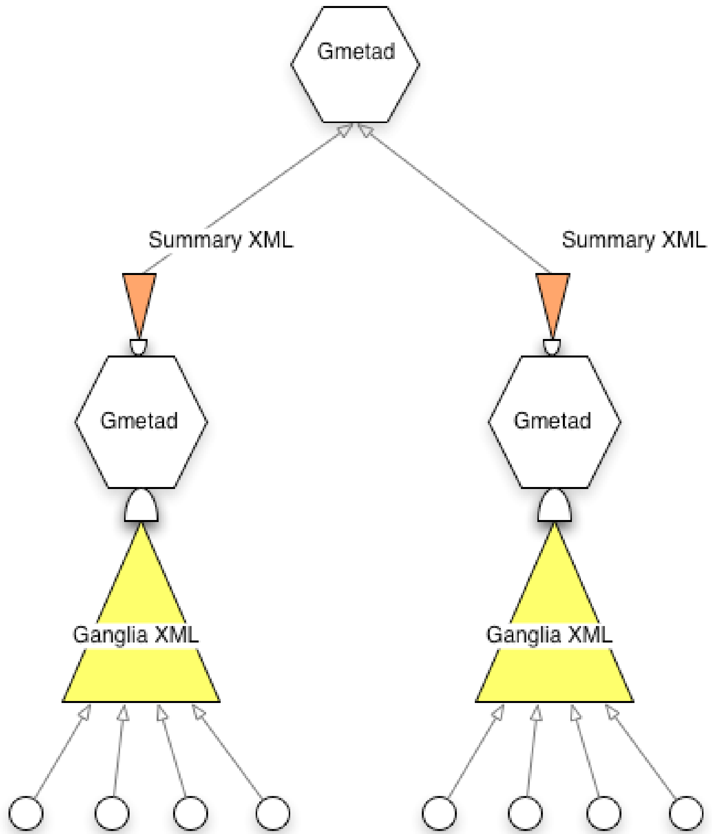
Monitoring Tree

- Experimental evidence: A single gmetad can only monitor ~1000 nodes before it is overwhelmed.
  - RRD metric history database maintenance is trouble spot.
- Solution: Tree structure.
  - Requires nodes to perform a data reduction.



Monitoring Tree

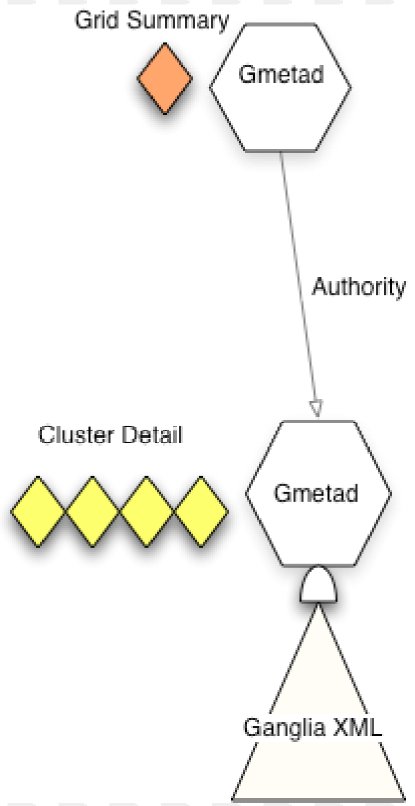
- Solution: Tree structure.
  - Requires nodes to perform a data reduction.
- Each gmetad does an average of its metrics
  - Only works for numeric metrics.
  - Provide sum, set size for each reduction.
- Enables *Summary* graphs



Ganglia Cluster Toolkit
http://ganglia.sourceforge.net

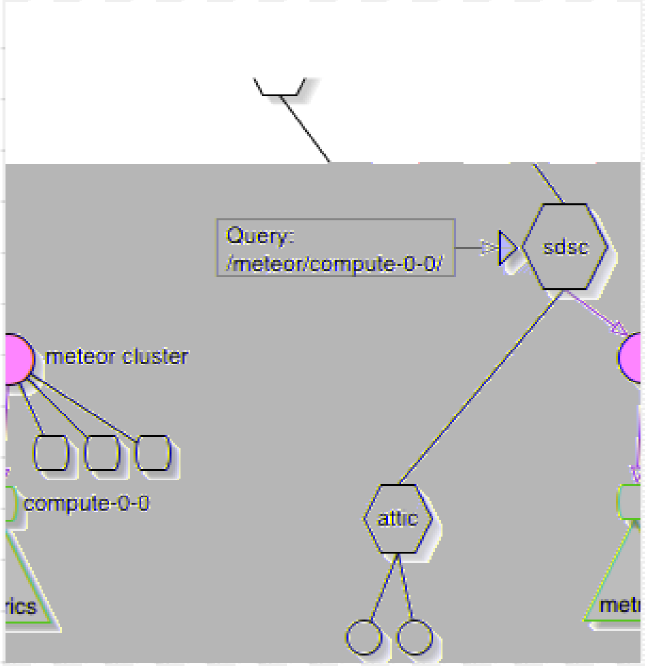
Delegation

- If we only have a summary, where is the raw data?
- Each summary metric graph is associated with an *authority pointer* (url) to a delegate gmetad.
- More detail is available from gmetads closer to the desired source in the monitoring tree.
  - Delegating work within tree is the main source of resource savings in our design
  - Less redundancy in system, in terms of metric history RRDs.



Query Support

- Ganglia web frontend: each page is generated using XML output from a local gmetad.
  - Queries are in the critical path
- Would like to return only relevant data for host views, etc.
  - XML parsing is slow.
  - Less stress on webserver
- New gmetad design introduces support for simple subtree queries.
  - XPath like
  - Implemented with Hash tables for O(1) lookup speed.



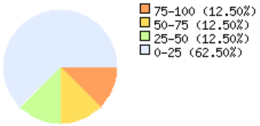
Query Support

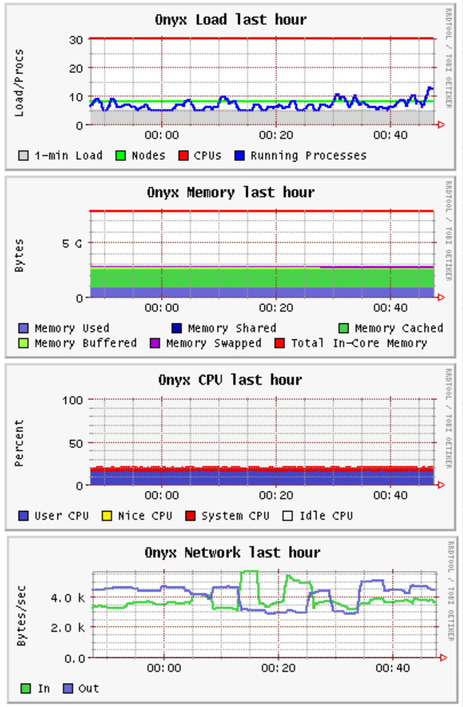
- Cluster view, Host View.
  - Query support makes host view more efficient.

CPUs Total:	**30**
Hosts up:	**8**
Hosts down:	**0**

Avg Load (15, 5, 1m):
17%, 17%, 17%
Localtime:
2003-11-24 00:47

Cluster Load Percentages





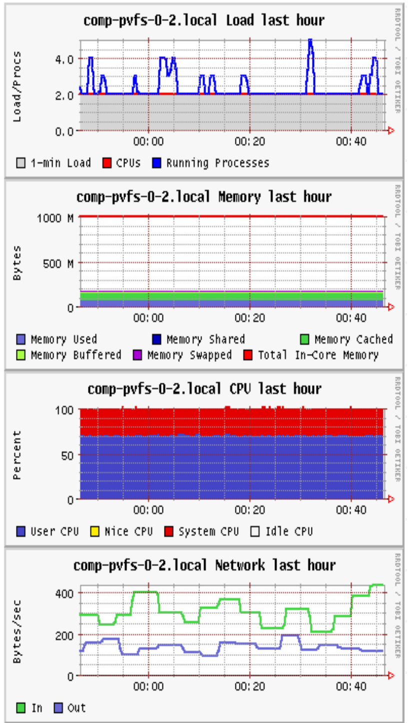
This host is up and running.

Time and String Metrics

Name	Value
boottime	Fri, 26 Sep 2003 17:38:24 +0000
gexec	OFF
machine_type	x86
os_name	Linux
os_release	2.4.20-20.7smp
sys_clock	Fri, 17 Oct 2003 20:20:21 +0000
uptime	58 days, 7:8

Constant Metrics

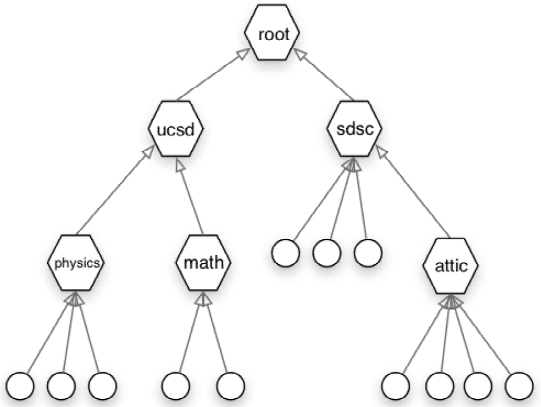
Name	Value
cpu_num	2
cpu_speed	2193 MHz
mem_total	1030596 KB
mtu	1500 B
swap_total	1020116 KB



Ganglia Cluster Toolkit
http://ganglia.sourceforge.net

Experimental Setup

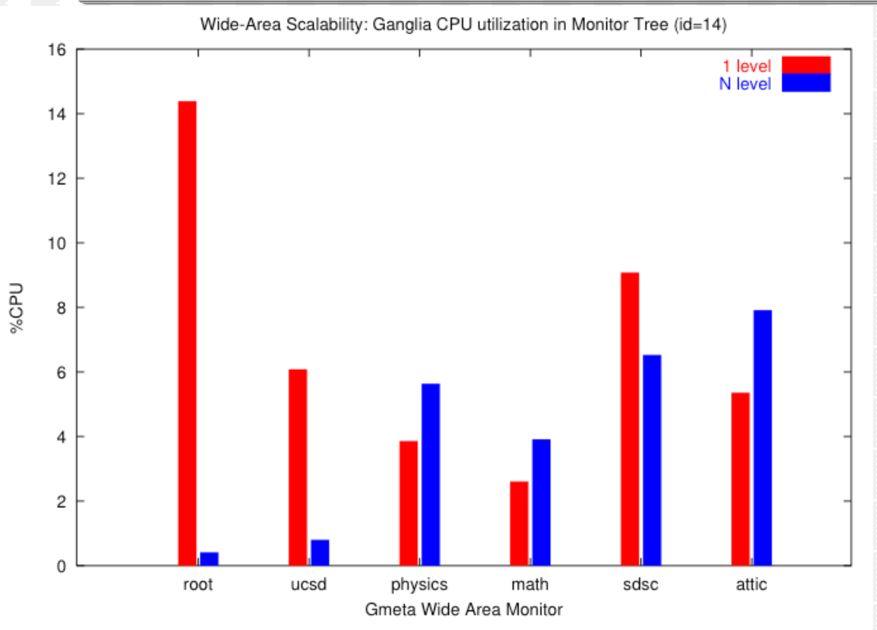
- Experiments to demonstrate the effectiveness of our monitoring technique.
  - 10-Node Linux cluster. Each node is 2-way P4 Xeon, 2.2Ghz, 1G memory.
  - Removed unnecessary variables
    - Disk I/O
    - Network bandwidth/latency to monitored clusters
  - 12 monitored clusters of identical size

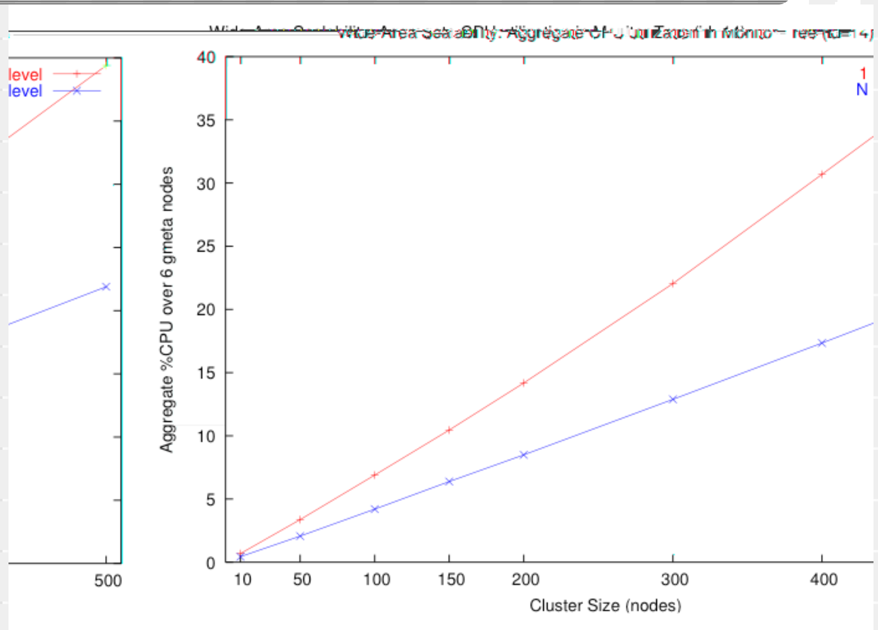


- Measured Average CPU utilization over 60 min of monitoring
  - Each gmetad runs on an otherwise unloaded cluster node

- *1-level* graph set is older Gmetad design: 2.5.1
- *N-level* is new Gmetad 2.5.4 using techniques from paper.

Experimental Results





- Load Delegation: CPU usage per gmetad node in monitoring tree.
- Tree monitors 12 clusters of 100 nodes each.

- Scaling: CPU usage sum over all gmetads for various cluster sizes.

Ganglia Cluster Toolkit
http://ganglia.sourceforge.net

Discussion

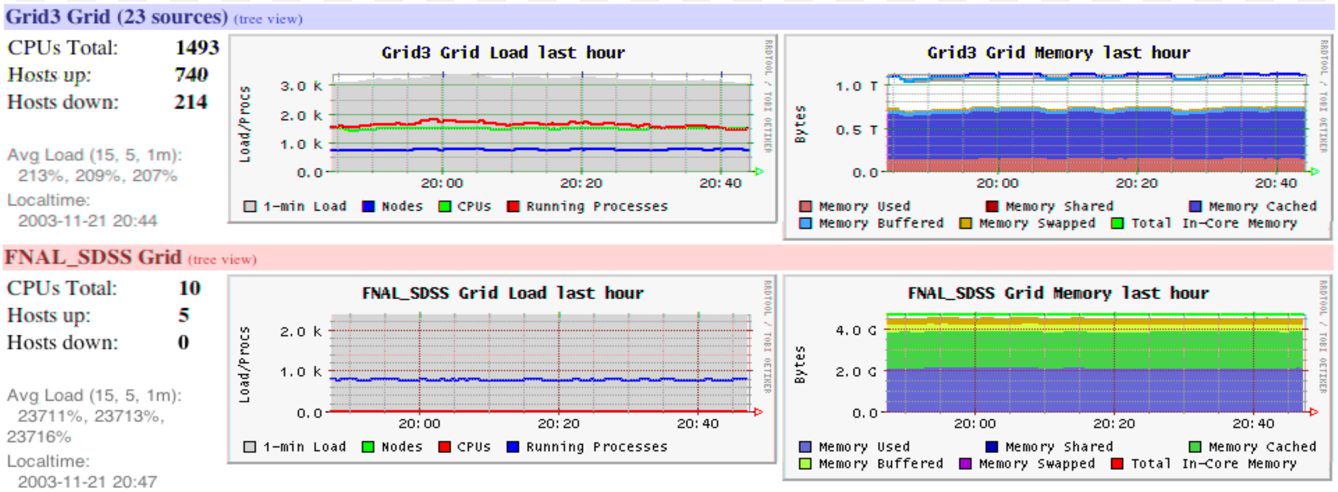
- Monitoring load moved away from root gmetad, transferred towards the leaves of tree *(physics, attic)*.
  - Shows scalability of design
- Leaf monitors (those closest to raw clusters) pay penalty for summarization tasks.
  - Higher load for leaf monitors, but acceptable.
- No bottleneck at root node as in previous design.
- Large speedup in webpage generation from query support as expected.

Limitations

- Trust model between Gmetad nodes in tree.
  - Must explicitly allow connections on both ends of edges.
    - MDS has a better trust model for essentially the same problem using Public Keys.
- No automatic failover between Gmetad monitors in the wide area.
  - Related to trust problem.
  - Difficult

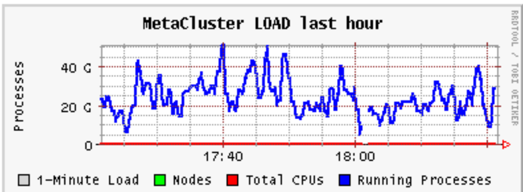
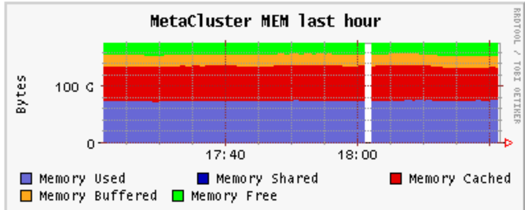
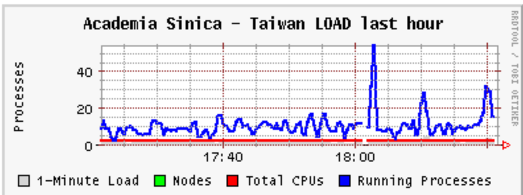
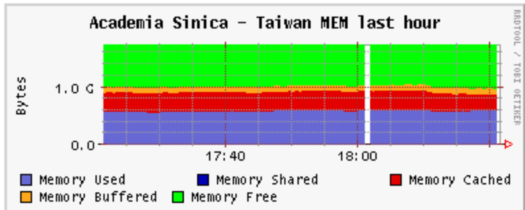
Ganglia Demos

- iVDGL Physics Grid
  - 800+ nodes worldwide, use new Ganglia with described techniques to distribute load.
    - Been stable for 6+ months at this scale.
  - http://gocmon.uits.iupui.edu/ganglia-webfrontend/



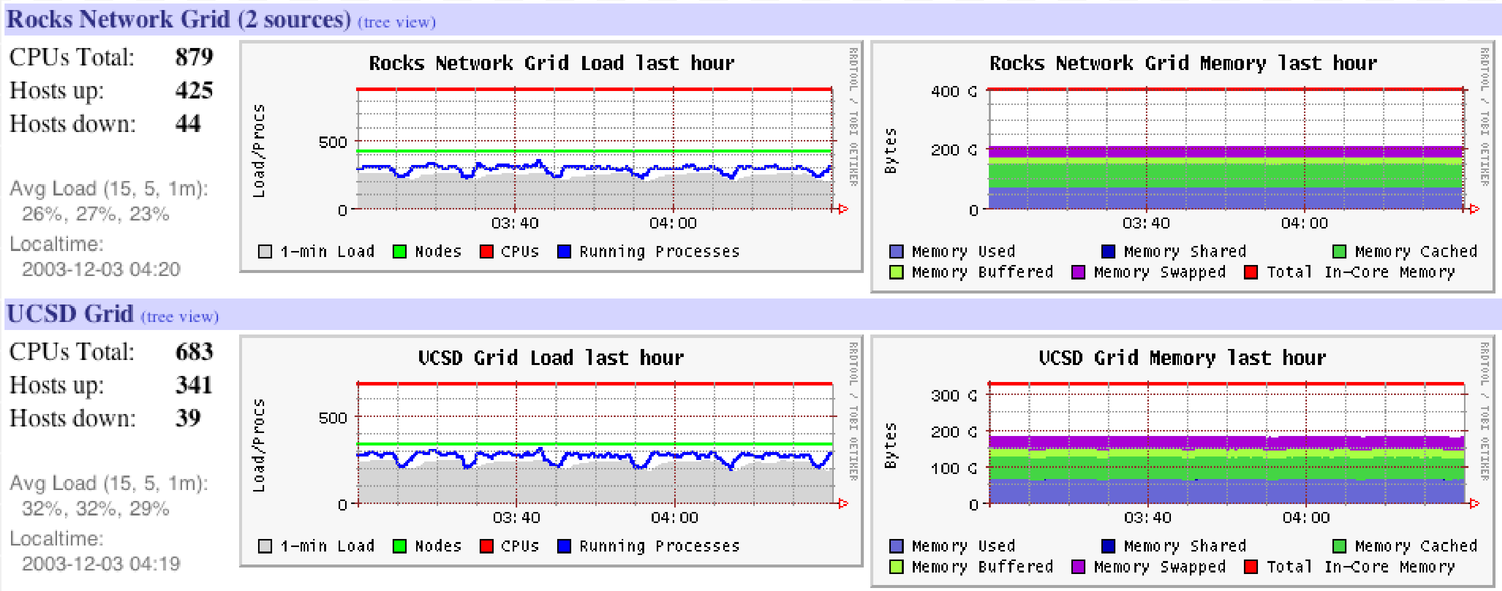
Ganglia Demos

- PlanetLab
  - Oldest Ganglia Grid: 93 clusters worldwide.
  - http://www.planet-lab.org/ganglia.beta/

Name / Info	Load Averages			%CPU User, Nice, System, Idle			
MetaCluster (Overall Data for 94 Clusters)	**2538.90**	**2552.58**	**4131.09**	**14.0**	**4.0**	**14.4**	**67.2**
224 hosts up and running (**224** CPUs Total) **7** hosts down							
Academia Sinica - Taiwan (physical view)	**2.53**	**2.16**	**1.68**	**36.0**	**0.8**	**63.8**	**0.0**
Cluster Localtime: **November 21, 2003, 6:21 pm** **2** hosts up and running (**2** CPUs Total) **0** hosts down							

Ganglia Demos

- Rocks Monitoring Network
  - 450 nodes, regional grid. Uses newest Ganglia design.
  - http://meta.rocksclusters.org/Rocks-Network/



Conclusion

- Technique presented that enables scalable cluster monitoring over the wide area.
  - Delegation model with additive reductions of data at every node.
  - Automatic, pointer-based method of navigating tree.
- Experiments show validity of design.
- Used in real-world projects on many different scales with good results.
- Shortcomings include interesting areas of future work.

Authors

- Federico Sacerdoti, SDSC (NPACI)
  - Fds@sdsc.edu
- Matt Massie, Berkeley (NPACI). Original Ganglia author, lead developer.
  - Massie@cs.berkeley.edu
- Mason Katz, SDSC. Cluster/Grid group leader
  - Mjk@sdsc.edu
- Dr. David Culler, Berkeley.
  - Culler@cs.berkeley.edu

Ganglia available at: Ganglia.sourceforge.net

Ganglia Cluster Toolkit http://ganglia.sourceforge.net