Ganglia

Ganglia: A Distributed Monitoring System

- **History**
- **Design**
- **Implementation**
- **Demonstration**

Origin of Ganglia

Ganglia is one of many software tools which grew out of the UCBerkeley Millennium Project. Millennium linked together small 16-CPU departmental clusters across the Berkeley campus to a large 320-CPU central cluster in Computer Science.

This ~$15 million computational infrastructure was created through a grant from the National Science Foundation with matching funds from Intel, Myricom, Nortel and Microsoft.

Ganglia History

Spring 1999: Brent Chun's rEXEC system
- used multicast for node discovery and state

Fall 1999: my ugly prototype
- perl daemons parsed `/proc` and wrote to MySQL

Spring 2000: my less ugly prototype (Ganglia 1.x)
- perl daemons, called *dendrites*, multicast state
- *Axons* then aggregated the multicast data in-memory and shared it via socket connections
- Web interface plugged into *axons* to get data
- Demo'd at SC2000 at the NPACI booth

Ganglia History Cont'd

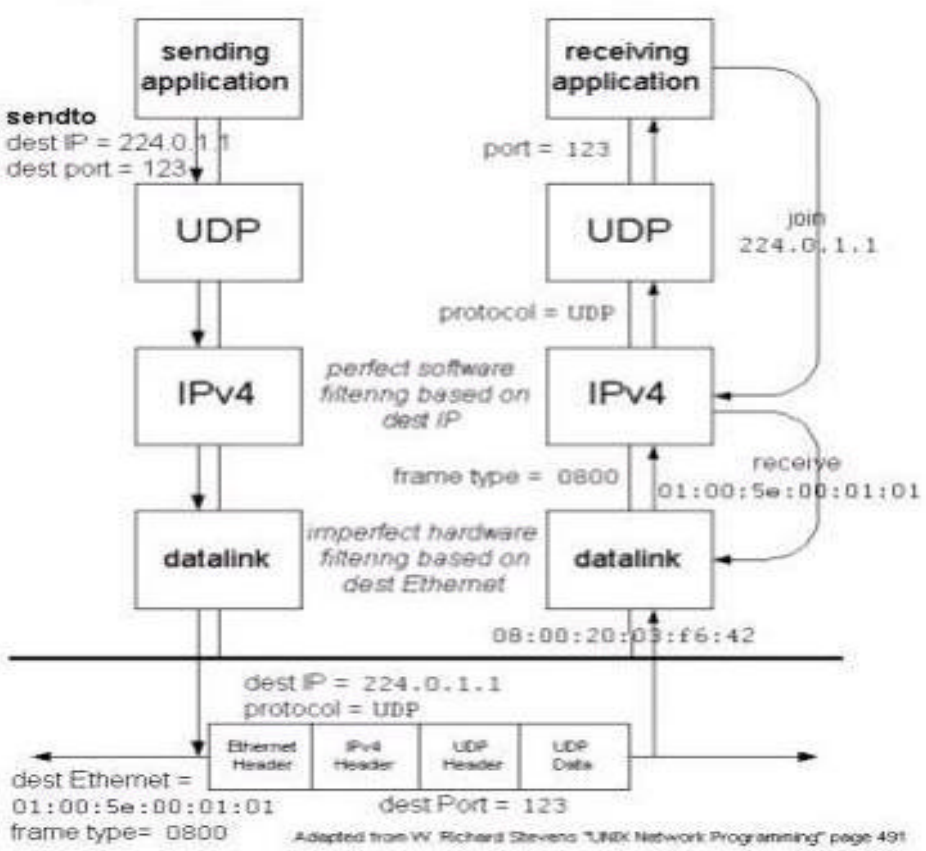
Spring 2001: Ganglia 2.x

- written completely in C
- dendrites and axons merged into gmond
- peer-to-peer model using XDR and XML

January 2002: Ganglia release on SourceForge

- ported to Linux (i386, ia64, sparc, alpha, powerpc, m68k, mips, arm, hppa, s390), Solaris, FreeBSD, AIX, IRIX, Tru64, HPUX, MacOS X (Darwin) and Windows (beta) and has been downloaded over 18,000 times

Ganglia Multicast



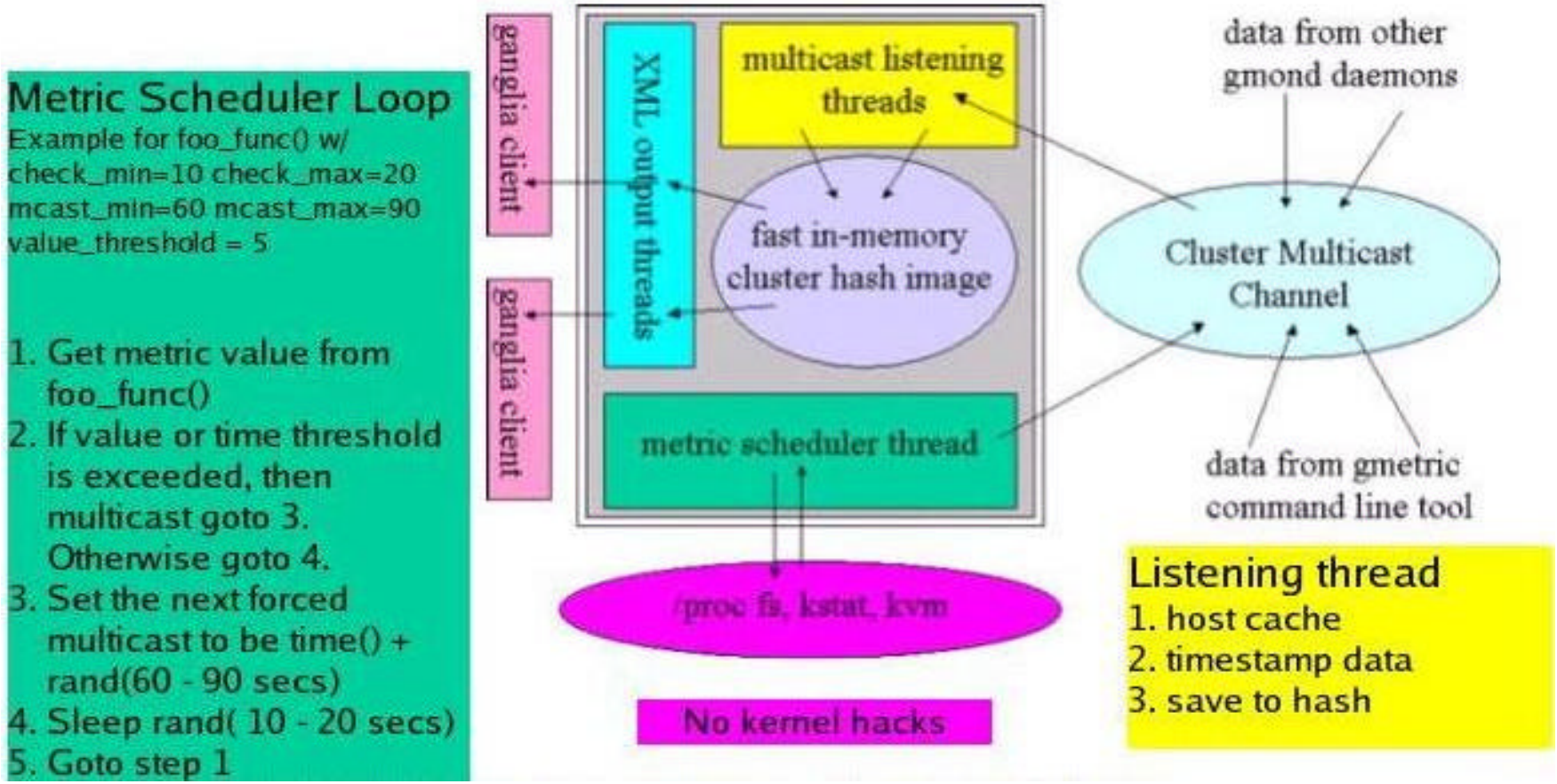
Advantages of Multicast

- Datagram received only by the interfaces that want it (making it great for node discovery)
- Group addresses
- Messages are UDP so there is no connection setup or teardown
- Can be used across a WAN

Disadvantages

- Not good for bulk data transfer (no windowed flow control, congestion avoidance, etc)
- UDP is not reliable

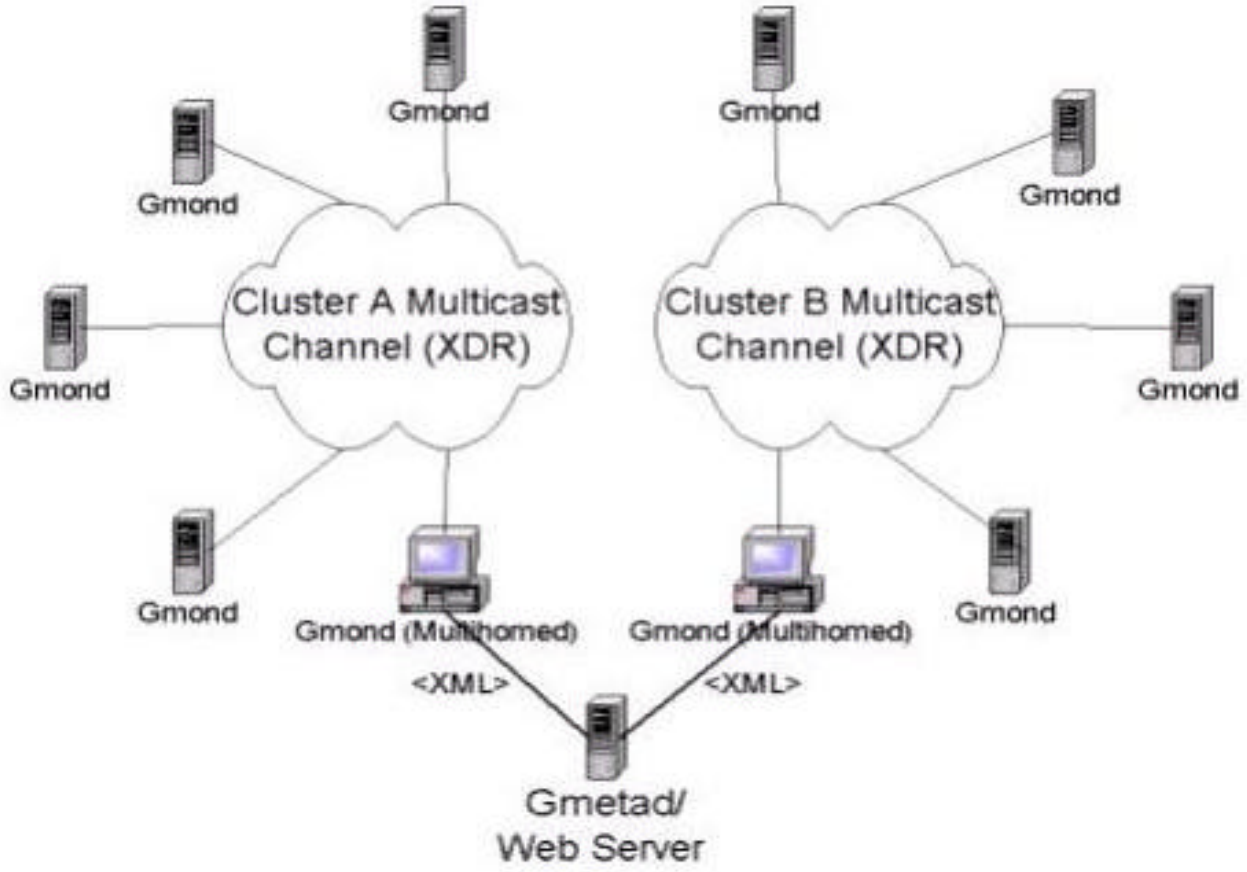
Monitoring Daemon (gmond)



gmond Metrics

Metric	Check Min	Check Max	Mcast Min	Mcast Max	Threshold
cpu_num	-1	-1	900	1200	-1
cpu_speed	-1	-1	900	1200	-1
mem_total	-1	-1	900	1200	-1
swap_total	-1	-1	900	1200	-1
boottime	-1	-1	900	1200	-1
sys_clock	-1	-1	900	1200	-1
machine_type	-1	-1	900	1200	-1
os_name	-1	-1	900	1200	-1
os_version	-1	-1	900	1200	-1
cpu_user	15	20	60	90	1
cpu_nice	15	20	60	90	1
cpu_system	15	20	60	90	1
cpu_idle	15	20	60	90	5
cpu_aidle	850	950	3400	3800	5
load_one	15	20	50	70	1
load_five	30	40	275	325	1
load_fifteen	60	80	850	950	1
proc_run	45	80	850	950	1
proc_total	45	80	850	950	1
mem_free	30	40	120	180	1024
mem_shared	30	40	120	180	1024
mem_buffers	30	40	120	180	1024
mem_cached	30	40	120	180	1024
swap_free	30	40	120	180	1024

Cluster Layout



Meta Daemon (gmetad)

- Multiple hosts per data source for redundancy
- Data sources have unique threads
- SAX XML parser with gperf hashes
- Exports a concatenated XML image of all clusters
- Data sources: either gmond or gmetad
  (gmetad will soon allow nesting of data sources)
- Historical data is saved to Round-Robin databases
  - Each rrdb is < 25KB
  - 1 year * 1000 nodes * 25 metrics < 300 MB

Example of XML

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
...(DTD)...
<GANGLIA_XML VERSION="2.5.0" SOURCE="gmond">
<CLUSTER NAME="My Cluster" LOCALTIME="1036308403" OWNER="unspecified" LATLONG="unspecified" URL="unspecified">
<HOST NAME="machine.Millennium.Berkeley.EDU" IP="169.229.50.111" REPORTED="1036308392" TN="11" TMAX="20" DMAX="0"
LOCATION="unspecified" GMOND_STARTED="1036109977">
<METRIC NAME="load_one" VAL="0.00" TYPE="float" UNITS="" TN="31" TMAX="70" DMAX="0" SLOPE="both" SOURCE="gmond"/>
<METRIC NAME="load_fifteen" VAL="0.00" TYPE="float" UNITS="" TN="262" TMAX="950" DMAX="0" SLOPE="both" SOURCE="gmond"/>
<METRIC NAME="cpu_nice" VAL="0.0" TYPE="float" UNITS="%" TN="9" TMAX="90" DMAX="0" SLOPE="both" SOURCE="gmond"/>
<METRIC NAME="cpu_user" VAL="0.0" TYPE="float" UNITS="%" TN="15" TMAX="90" DMAX="0" SLOPE="both" SOURCE="gmond"/>
<METRIC NAME="proc_total" VAL="71" TYPE="uint32" UNITS="" TN="259" TMAX="950" DMAX="0" SLOPE="both" SOURCE="gmond"/>
...
<METRIC NAME="cpu_idle" VAL="99.9" TYPE="float" UNITS="%" TN="57" TMAX="90" DMAX="0" SLOPE="both" SOURCE="gmond"/>
<METRIC NAME="load_five" VAL="0.00" TYPE="float" UNITS="" TN="44" TMAX="325" DMAX="0" SLOPE="both" SOURCE="gmond"/>
<METRIC NAME="cpu_num" VAL="1" TYPE="uint16" UNITS="" TN="308" TMAX="1200" DMAX="0" SLOPE="zero" SOURCE="gmond"/>
<METRIC NAME="disk_free" VAL="1.109" TYPE="double" UNITS="GB" TN="90" TMAX="180" DMAX="0" SLOPE="both"
SOURCE="gmond"/>
<METRIC NAME="os_name" VAL="Linux" TYPE="string" UNITS="" TN="320" TMAX="1200" DMAX="0" SLOPE="zero"
SOURCE="gmond"/>
</HOST>
</CLUSTER>
</GANGLIA_XML>
```

Mason Katz and Carl Kesselman created software to translate ganglia XML into Glue Schema LDIF to allow ganglia data to be exported to the Globus MDS

Web Frontend

Web Frontend

- Maintained by Federico Sacerdoti
- TemplatePower separates PHP code from HTML
- Requires PHP-enabled webserver with local Ganglia Meta Daemon (gmetad)
- Three levels of presentation: grid, cluster and host
- Will soon use a hierarchical delegation model very similar to DNS

Scalability

The PlanetLab open testbed is currently using a stock 2.5.0 distribution of ganglia to monitoring the following 42 sites across 5 different countries in real time using a single gmetad:

Arizona, CalTech, Cambridge (U.K.), Carnegie Mellon, Columbia, U of Copenhagen (Denmark), Cornell, Duke, GT, Harvard, ISI, Intel Research (Oregon, Berkeley, Pittsburgh, Seattle), U of Kansas, U of Kentucky, LBL, Lancaster (U.K.), MIT, Michigan, NYU, Princeton, Rice, Stanford, Tennessee, UBC, UC (Berkeley, San Diego, Santa Barbara, LA), UMass, UPenn, U of Canterbury (New Zealand), Univ of Technology, Sydney (Australia), U of Texas Austin, University of Bologna (Italy), U of Washington, Uppsala U (Sweden), Utah, WashU, and Wisconsin

The gmetad is running on a 1.25 GHz 1GB Linux box. Gmetad uses about 93 MBs of memory and never more than 2% CPU.

Visit http://www.planet-lab.org/ click "Real-time status" to see ganglia report

Future (Ganglia 3)

- Hierarchical data model both in the internal data respresentation and the XML (no more flat namespace)
- Hierarchical Dynamic Shared Object (DSO) plugins which take responsibility for a particular branch of the overall tree
- Gmond/Gmetad data sources will be interactive allowing a subset of the overall data space to be explored
- Trigger plugins will allow handling of events in near-real time
- A rich ganglia query system will be available based on an open-source XML Query engine (Qexo, Kweelt, XQuench)
- The gmetad and web frontend will use a hierarchical delegation model very similar to DNS (new XML tags)
- XML data compression

Developers

Brent Chun (Intel Research at Berkeley)
Ganglia Execution System
Preston Smith (Purdue University Physics)
AIX Port, FreeBSD Port, MacOS X Port, Linux cleanup
Steve Wagner
IRIX Port, Tru64 Port, Solaris Port
Federico Sacerdoti (San Diego Supercomputer Center)
Web Frontend, Monitor Core Features/Bugfixes, PBS Integration
Greg Bruno (San Diego Supercomputer Center)
Ganglia Pyton Class and Client
Mason Katz (San Diego Supercomputer Center)
Ganglia Python Class, Globus Integration
Matt Rice
Windows (Cygwin) Port
Martin Knoblauch
HPUX Port
Alan Hagge
Preliminary IRIX Support
Dave Wallace, Chris Elmquist, Asaph Zemach, Leif Nixon, Jeff Layton, Doc Schneider, Oliver Mössinger, Tim Cera, Mathew Benson, Davide Tachella, Joshua J England, Yaroslav Klyukin, Neil Spring, Brian Youngstrom
Submitted important feature and bugfix patches

Ganglia Users

- NASA Goddard Space Flight Center
- National Research Laboratory of the US Navy
- Centre National De La Recherche Scientifique
- San Diego Supercomputing Center
- Intel
- ArrayNetworks
- BellSouth Internet Services
- Cray
- Dow Chemical
- Lockheed Martin
- nVidia
- Dell
- Boeing
- Sandia National Laboratory
- Keldysh Institute of Applied Mathematics (Russia)
- The Max Planck Society for the Advancement of Science (Germany)
- Optiglobe Brazil (Brazil)
- Brunel University (U.K.)