A Step Towards Automated Event Diagnosis

Stanford Linear Accelerator Center

IEPM

Internet End-to-end Performance Monitoring

Adnan Iqbal, Yee-Ting Li, Les Cottrell

Connie A. Log. Williams Jerrod

In this presentation

- Cause of Problems
- Background
- Motivation
- Basic Principles
- Case Study
- Limitations
- Current Status

Cause of problems

- Hardware
  - Network Device failures
  - Communication link failure
  - Fully loaded machines
- Software/Configuration
  - PC configuration
  - Badly designed/configured applications
  - Routing inconsistencies
- Maintenance/Upgrade
- Firewalls

Cause of problems

- Network Device failures
  - Stanford Connectivity Failure (Sat) 10th September 2005 ~14:00
    - CENIC had an unplanned outage of the Stanford 15540 due to a fan failure.
    - Effected SLAC's paths to CENIC (and Internet 2)
- Communication link failure
  - BNL August 2005
    - Event reported from BNL to several sites at 7:45pm
    - BNL's primary ESNet fiber connection (OC-48) to NYC went down on 8/16/05 at around 6:45pm EDT.
    - At that time BNL's only connection to the internet was through its secondary backup connection (T3) through NYSernet. The primary link was restored at around 3:44am EDT.
  - Pakistan July 2005
    - Under water fiber cable (SEAMEWE-III) damaged.
    - Fault was corrected after eight days.

Cause of problems

- Maintenance/Upgrade
  - April 2006 CERN-TENET
    - Anomalously large min-RTT reported between 9 and 10 April 2006
    - TENET shifted to Abilene without prior warning
- Badly designed/configured applications
  - SLAC-CALTECH applet problem
    - November 2005, multiple alerts reported from CALTECH to SLAC and SLAC to CALTECH
    - An applet running on both ends was causing problem, after killing it from both ends every thing was back to normal.
    - The application was opening sockets for communication but not closing them.

Cause of problem

- Routing protocol inconsistencies
  - SLAC to CALTECH factor of 5 drop in performance
    - iperf throughput drop reported on August 27, 2003.
    - a CENIC router in Los Angeles (ASN 2152) was receiving Caltech's prefixes via a Los Nettos route server on a shared connector segment.
    - The Los Nettos route server was preferring paths to Caltech that went through a next hop that was not reachable from the CENIC router (and was then advertising that next-hop to the CENIC router).
    - Because of the unreachable next-hop the CENIC router was re-writing the advertised next-hop to be the direct peering address of the Los Nettos route server.
    - The los nettos route server's unreachable-from-CENIC path traverses a 100 Mb/s ethernet. This was the cause of the bottleneck.
    - Manual change of route, corrected the problem

Motivation

- All these problems causes detectable changes in measured metric values
  - Experienced connectivity, latency, jitter, network path, throughput, user expectations
  - An event is when one/many anomaly/ies occur
  - So infer that when an event occurs, a problem has been detected!
- Analyze events for
  - Event Isolation (finding cause of events)
    - Replace old hardware, reconfiguration, change peerings
  - Event Relationship (between different events)
    - Cascade effects, check backup solution is working
  - Confirmation of events (False positives)
    - Seasonal effects etc. planned outages

Motivation

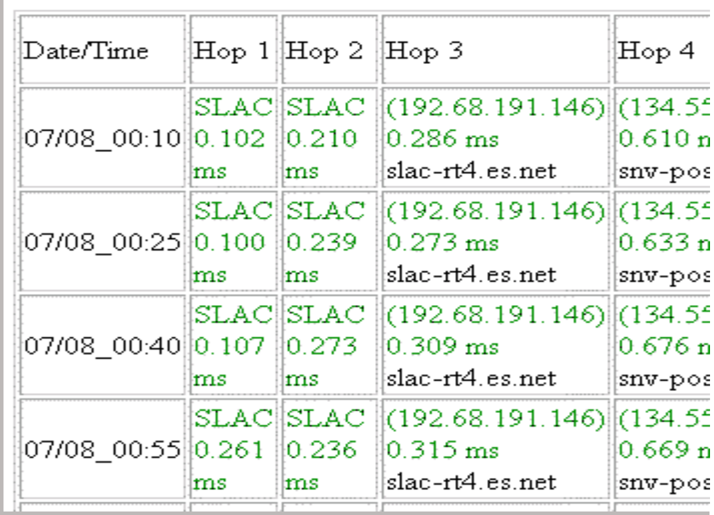
- Currently End-to-End Monitoring and Diagnosis very Laborious:
  - Looking at graphs of data
  - Searching for route changes
  - Reconciling data from many different databases manually
  - Occasionally consulting other tools/services
    - Nagios, perfSONAR, maintenance tickets, phone calls etc.
- Analysis is usually conducted late.
  - Have to infer problem scenario and guess at event cause post event
  - Cannot conduct further tests during problem to confirm cause(s)
- The process is ad hoc
  - does not follow any specific procedure so always a chance of missing some important information
  - Need to build a ‘logic database’ to aid diagnostics: heuristics

Goals

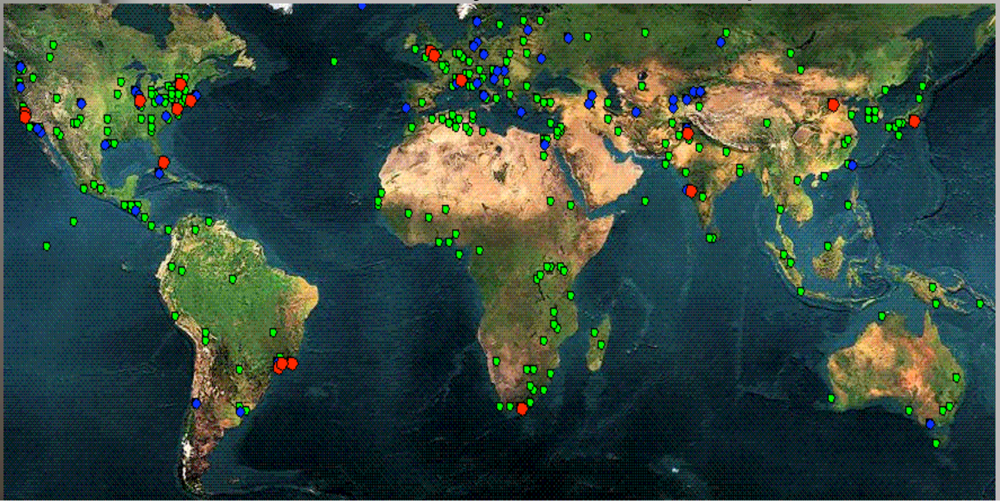
- We want to find cause of a reported event as soon as possible
- If we keep on doing it manually, we cannot do it quickly
- Automation is not easy
  - Nature of problems varies
  - An apparent cause may not be the actual cause
  - One unified technique may not be applicable
- So what is our approach?
  - Define heuristics which describe relationship between an apparent symptom (event) and a possible cause (problem)
  - Use these heuristics to find out actual cause of problem
  - These rules may be complex.
    - One problem may lead to many events
    - One event may be due to many problems
  - Use a simple scoring system to determine most probable cause

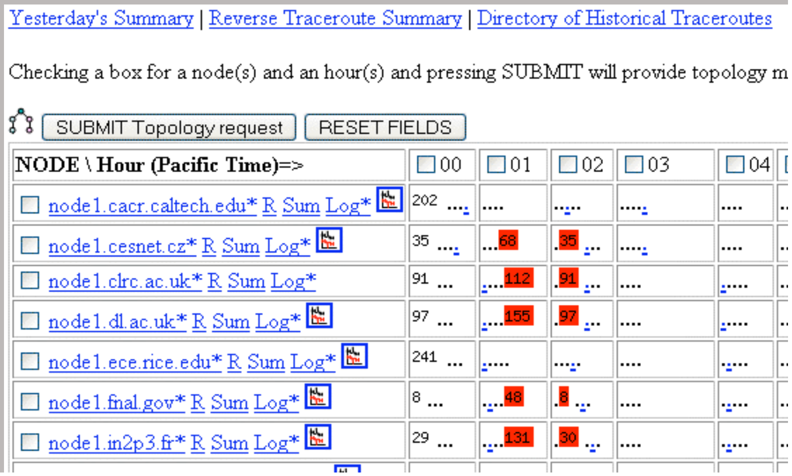
IEPM-BW Background

- IEPM-BW Deployment
  - Different Network Monitoring tools
    - Ping, IPerf, Thrulay, Pathchirp, Pathload…
  - Variety of Metrics
    - Throughput, RTT, available bandwidth
  - Traceroutes
- Topology
  - 5 Monitoring Hosts
  - Over 25 nodes being monitored by each host
  - Number of Monitoring Host and nodes both increasing
  - Currently, bulk is in Europe and North America



Date/Time	Hop 1	Hop 2	Hop 3	Hop 4
07/08_00:10	SLAC 0.102 ms	SLAC 0.210 ms	(192.68.191.146) 0.286 ms slac-rt4.es.net	(134.55 0.610 n snv-pos
07/08_00:25	SLAC 0.100 ms	SLAC 0.239 ms	(192.68.191.146) 0.273 ms slac-rt4.es.net	(134.55 0.633 n snv-pos
07/08_00:40	SLAC 0.107 ms	SLAC 0.273 ms	(192.68.191.146) 0.309 ms slac-rt4.es.net	(134.55 0.676 n snv-pos
07/08_00:55	SLAC 0.261 ms	SLAC 0.236 ms	(192.68.191.146) 0.315 ms slac-rt4.es.net	(134.55 0.669 n snv-pos





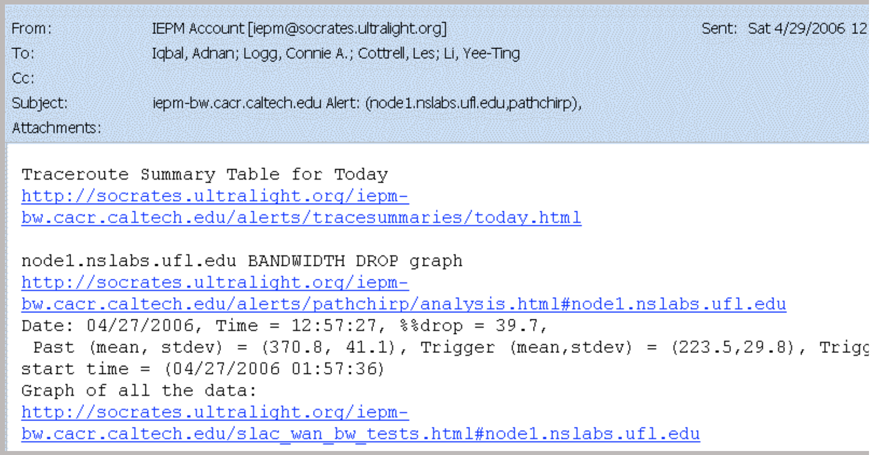
Yesterday's Summary | Reverse Traceroute Summary | Directory of Historical Traceroutes

Checking a box for a node(s) and an hour(s) and pressing SUBMIT will provide topology m

NODE \ Hour (Pacific Time)=>	□ 00	□ 01	□ 02	□ 03	□ 04
□ node1.cacr.caltech.edu* R Sum Log*	202 ....:	....	..:..	....:	....
□ node1.cesnet.cz* R Sum Log*	35 ....:	...68	.35 ..:.	....:	....
□ node1.clrc.ac.uk* R Sum Log*	91 ...	:...112	.91 ..:.	....	:....
□ node1.dl.ac.uk* R Sum Log*	97 ...	:...155	.97 ..:.	....	:....
□ node1.ece.rice.edu* R Sum Log*	241 ...	:....	...:.	....	.:...
□ node1.fnal.gov* R Sum Log*	8 ...	..:..48	.8 ..:.	....	.:...
□ node1.in2p3.fr* R Sum Log*	29 ...	..:..131	.30 ..:.	....	.:...

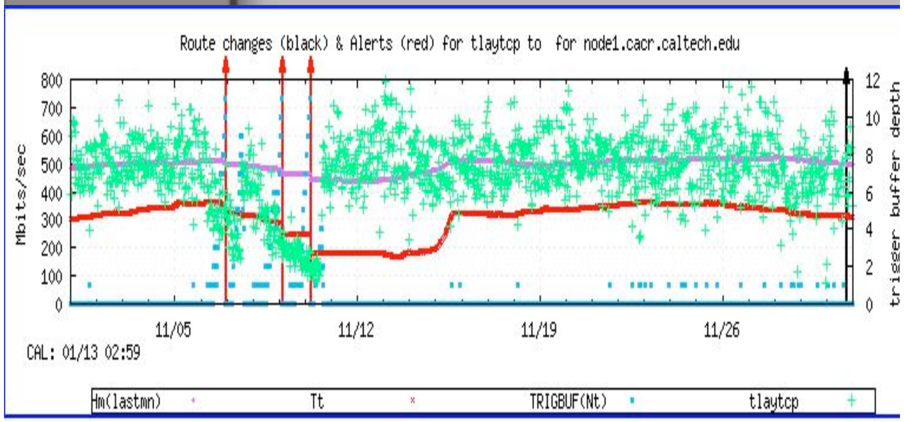
Automated Event Detection

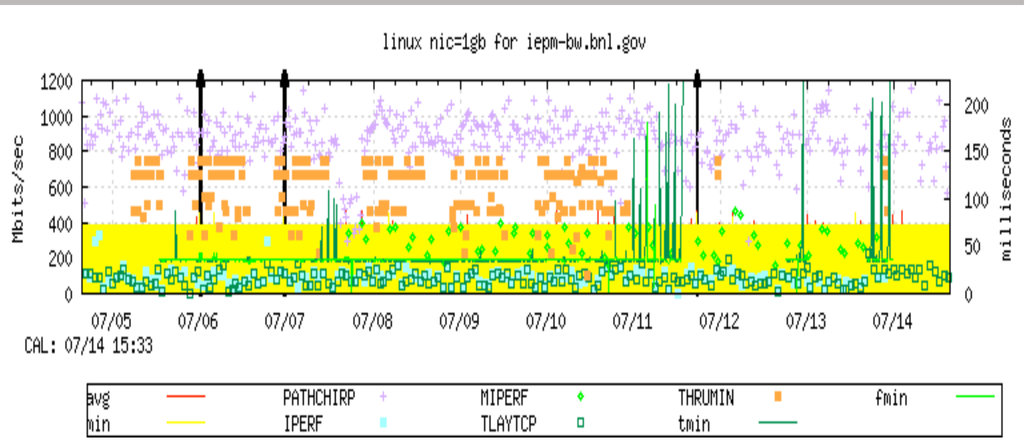
- Different Mechanisms
  - Holt-winters, Plateau
- Alerts are
  - Detected
  - Reported through e-mail
  - Stored in database
  - Made available on web



From: IEPM Account [iepm@socrates.ultralight.org] Sent: Sat 4/29/2006 12
To: Iqbal, Adnan; Logg, Connie A.; Cottrell, Les; Li, Yee-Ting
Cc:
Subject: iepm-bw.cacr.caltech.edu Alert: (node1.nslabs.ufl.edu,pathchirp),
Attachments:

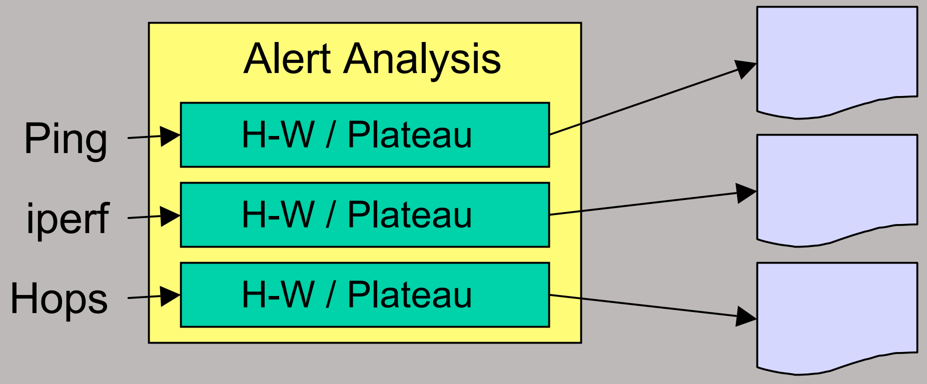
```
Traceroute Summary Table for Today
http://socrates.ultralight.org/iepm-
bw.cacr.caltech.edu/alerts/tracesummaries/today.html

node1.nslabs.ufl.edu BANDWIDTH DROP graph
http://socrates.ultralight.org/iepm-
bw.cacr.caltech.edu/alerts/pathchirp/analysis.html#node1.nslabs.ufl.edu
Date: 04/27/2006, Time = 12:57:27, %%drop = 39.7,
 Past (mean, stdev) = (370.8, 41.1), Trigger (mean,stdev) = (223.5,29.8), Trigg
start time = (04/27/2006 01:57:36)
Graph of all the data:
http://socrates.ultralight.org/iepm-
bw.cacr.caltech.edu/slac_wan_bw_tests.html#node1.nslabs.ufl.edu
```





IEPM-BW Event Detection Design Overview

- On each machine
  - Measurements taken…
  - Are analysed for events…
  - Stored on local machine in DB

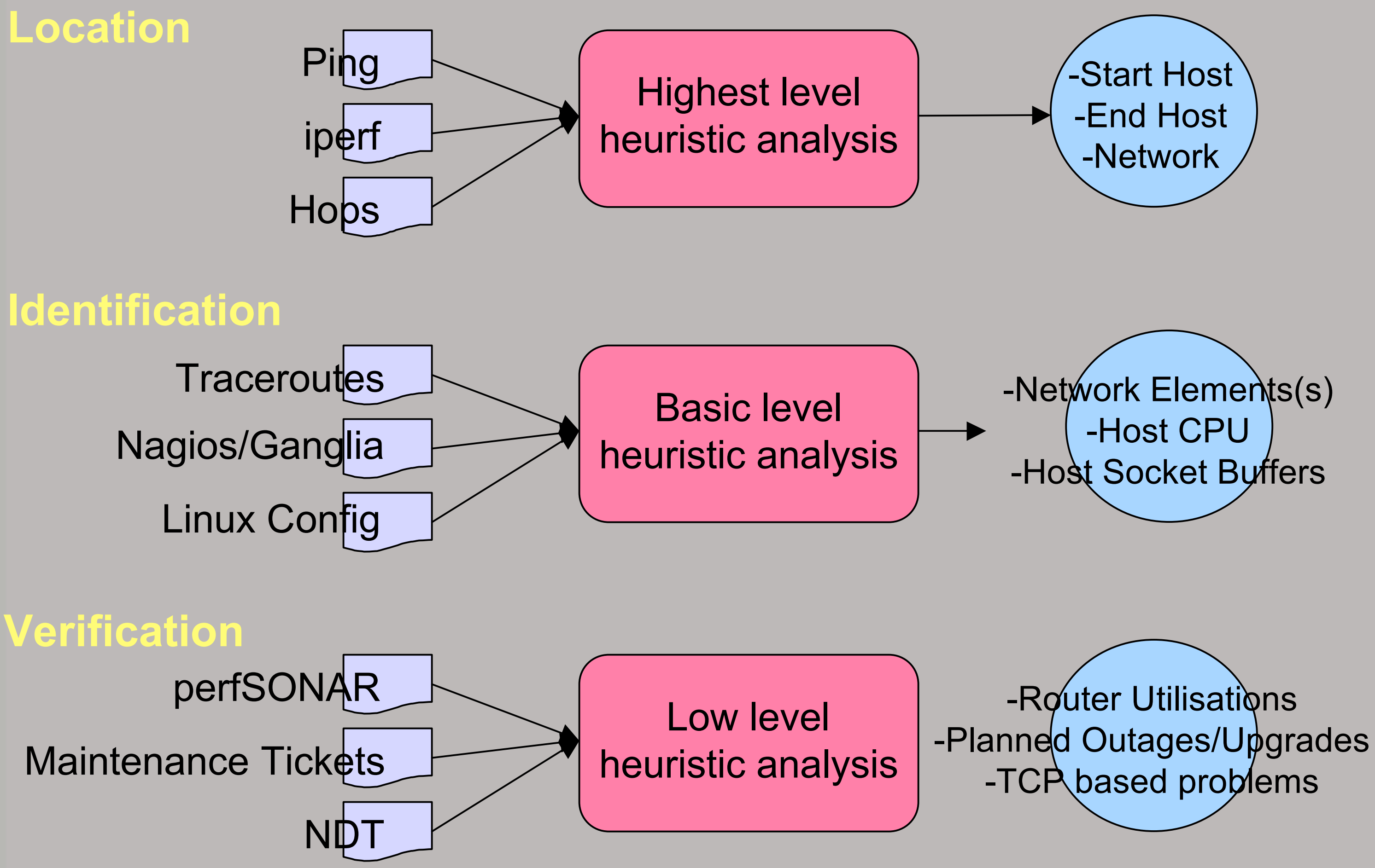


Event Diagnosis Overview

- End-to-end analysis; assume that the problem may not just be from the network
- Event initiated when a metric experiences an alert (anomaly through HW/Plateau).
- Initiate analysis of all available results
- Search Event DB for each metric checked on each host and used to cross correlate results between Monitor{ed|ing} Hosts
- Use heuristics defined with a simple summation algorithm to narrow down the location of a fault.
- Tiered system used to pin-point exact cause of problem.
  - Start from high level to very low level to help locate, identify and verify cause of alert

Detection

Analysis Overview



Location

Ping

iperf

Hops

Highest level heuristic analysis

-Start Host
-End Host
-Network

Identification

Traceroutes

Nagios/Ganglia

Linux Config

Basic level heuristic analysis

-Network Elements(s)
-Host CPU
-Host Socket Buffers

Verification

perfSONAR

Maintenance Tickets

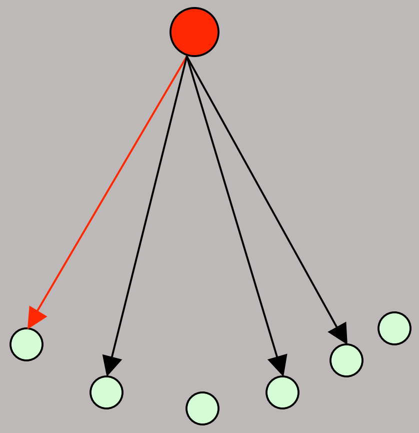
NDT

Low level heuristic analysis

-Router Utilisations
-Planned Outages/Upgrades
-TCP based problems

Heuristics

- If an event E is detected at time T, from monitoring host H for a monitored node M and there exist other events reported by same monitoring host H for monitored nodes other than M, than probability that monitoring host is causing problem increases with every such result



$$S_{smh} = \sum_{i=1}^{n} \alpha$$

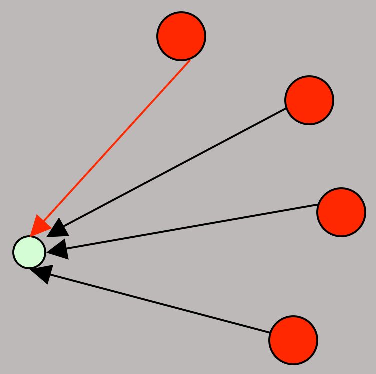
Where i represents only those nodes which reported event and α represents the score each such incident

Heuristics

- If an event E is detected at time T, from monitoring host H for a monitored node M and there exist other events reported by different monitoring hosts for same monitored node, than probability that monitoring node is causing problem increases with every such result

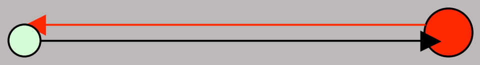
$$S_{node} = \sum_{i=1}^{n} \beta$$

Where i represents only those hosts which reported event and β represents the score each such incident



Heuristics

- If an event E is detected at time T, from monitoring host H for a monitored node M and there exist other events reported by M for H in similar time period than it confirms that event is not a false alarm. However, its cause can be any of the route change, network problem and or any of the end host.



$$S_{node} = \gamma$$

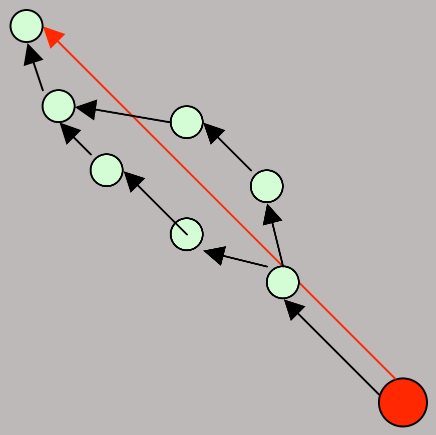
$$S_{host} = \mu$$

$$S_{network} = \lambda$$

Where S_x represents score for each category x.

Heuristics

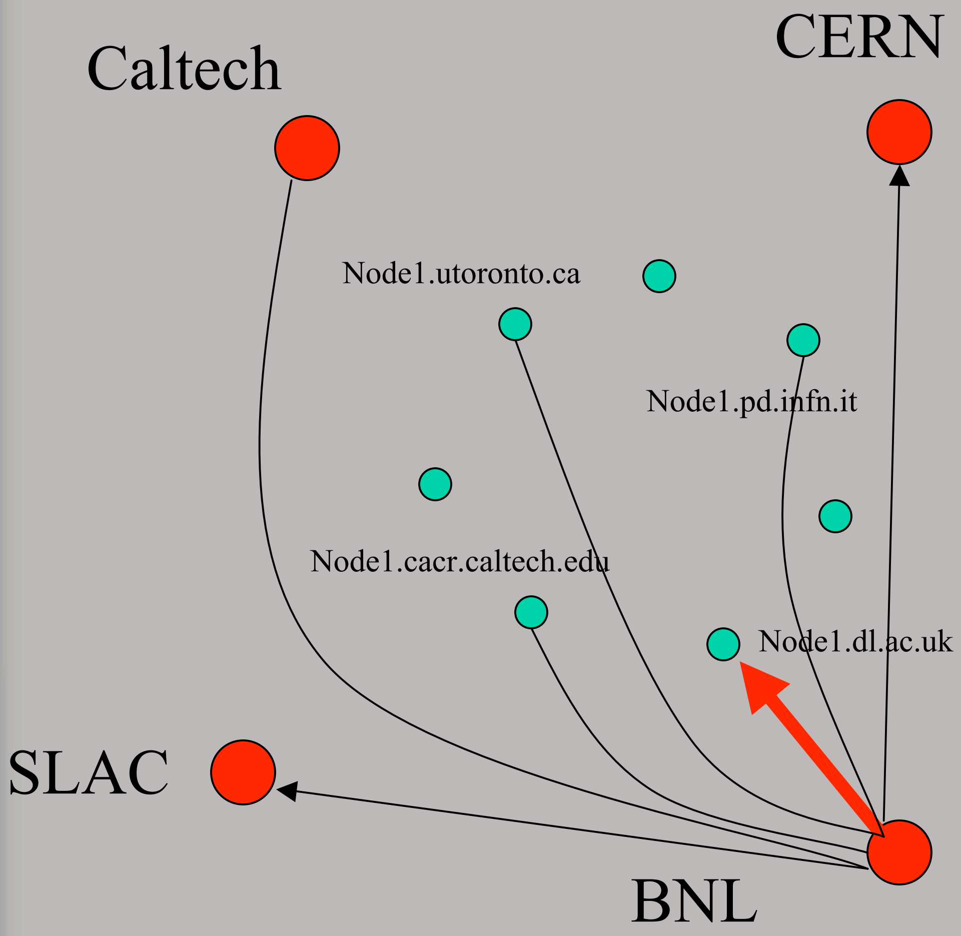
- If an event E is detected at time T, from monitoring host H for a monitored node M and there exist a change in the *number of hops* at the time of event detection than possibility is that performance drop is caused by any network problem.



$$S_{network} = \lambda$$

Where $S_{network}$ represents score for network problem.

BNL problem 12/30/05



Hop # Change No

Reverse Path does not exist

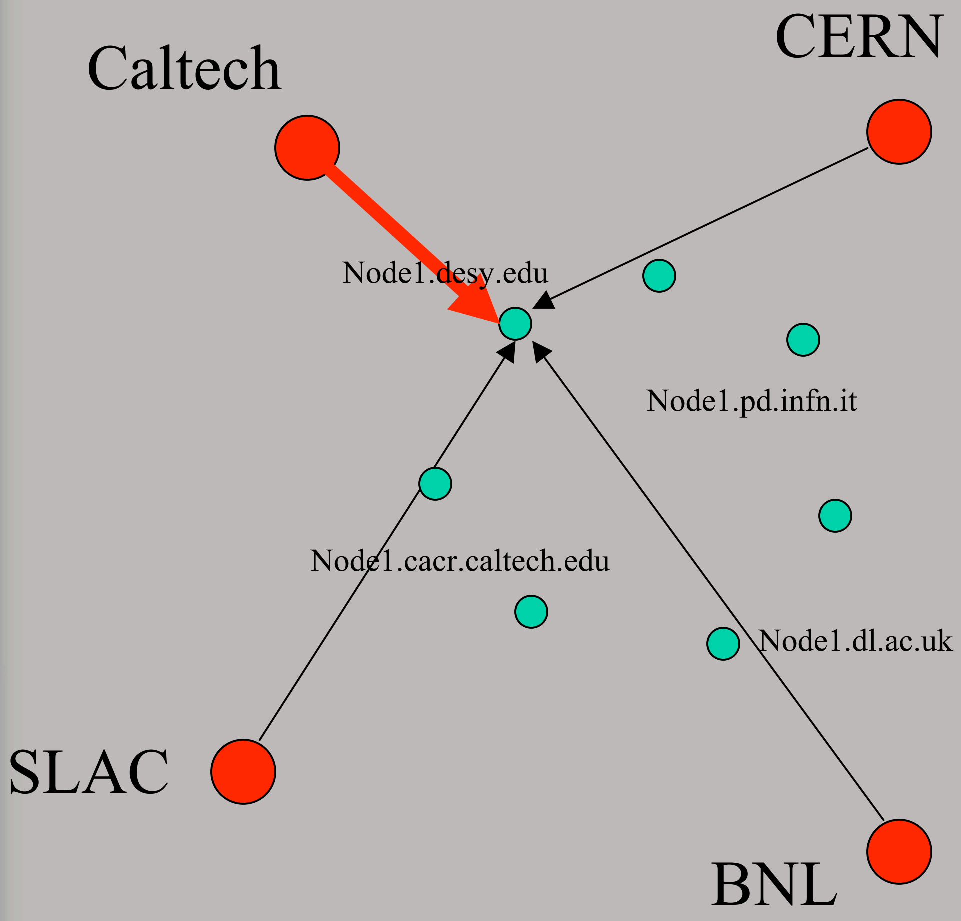
Other Hosts to DL no event reported

BNL to other Nodes Number of events reported

Possible Cause
Monitoring Host (BNL) is not working Fine.

http://www.slac.stanford.edu/grp/scs/net/case/bnl_dec05/

DESY problem 01/30/06



Hop # Change No

Reverse Path does not exist

Other Hosts to DESY every monitoring node reported an event and some reported multiple tools

Caltech to other Nodes No event reported

Possible Cause

Monitored Node (DESY) is not working correctly.

Limitations

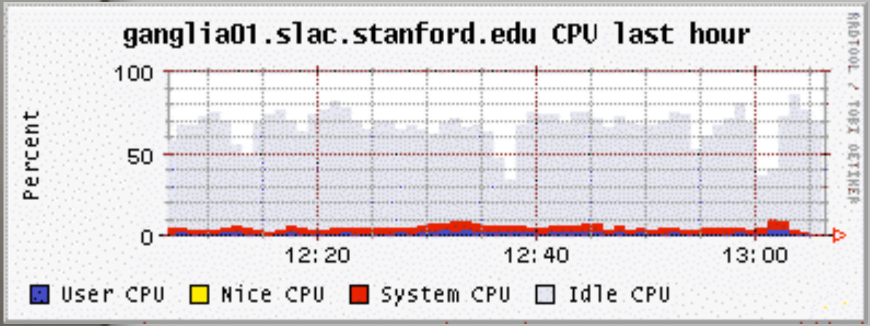
- Gives a good guess where the problem lies but does not confirm it 100%
  - Use further tests to isolate the identification of specific problem, and then verify the problem.
- To get a final statement, few more things are required
  - Tools that can provide some information about the condition of network at a previous given time
  - Tools that can provide statistics about end hosts at a previous given time

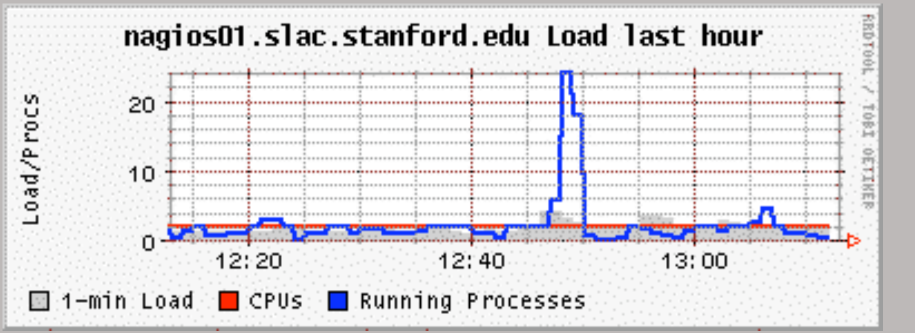
Network Problem Diagnosis

- perfSONAR
  - – Currently use IEPM-BW db - no reason it can’t use Network statistics from MA’s
  - – Event detection module for perfSONAR?
  - – The wider the deployment, the more data we can get, the more case studies we can analyse, the more helpful it is to pinpoint the problem(s)… the more advanced our heuristics can be
  - – Questions are what metrics would be useful?
    - Utilisation, Packet Drops,

End Host Problem Diagnosis

- Ganglia & Nagios
  - Number of end host statistics, easy installation and configuration
  - Web interface with number of graphs
  - http://ganglia01.slac.stanford.edu:8080/ganglia/monsystems/?r=hour&c=ganglia-monitoring&h=ganglia01.slac.stanford.edu
  - Need access to other site’s monitoring.





Current Status

- Includes
  - Initial information about alert
  - Results for each analysis
  - Related links
  - Trace route changes, details and AS traversed
  - Final scores
- Publicly available on web
  - http://www-iepm.slac.stanford.edu/monitoring/event-diagnosis/analysis/case1.html

Reports/Issues

- What other tools/metrics are available out there to generate these (extra) reports?
  - BER on a line
  - MIBs: consistency/conversation/unification
  - Data overload?!?
- What (extra) reports and logic are needed to help pin-point event causes further?
  - What do people/engineers *do* to diagnose problems?
  - Need community feedback and design
- Availability of Data for Diagnosis
  - Solution solved somewhat by perfSONAR (local control of data availability, AuthN/Z)
  - Eg Routing tables etc.
- Traceroutes
  - Transparency? Eg Layer 2, Light paths?
  - Machine (logical) naming of router ports
- Maintenance tickets
  - Machine parseable?

Summary

- Many different problems can lead to events
- Identify and categorise events to create heuristics
- Logic of heuristics used to diagnose why event occurred
- Used simple summation of heuristic metrics to determine most likely cause (network or host)
- Need more detailed heuristics to help really identify, validate and isolate problems!
  - Need access to metrics!
  - Need experience from experts!