Monitoring e2e Performance on High-speed Networks

Margaret Murray (CAIDA)

Acknowledgements

- Shava Smallen: TeraGrid INCA Test Harness Framework at SDSC
- Omid Khalili: INCA reporter programming
- Nevil Brownlee: NeTraMet development and config
- Johnny Chang, Alok Shriram: bwest tool evaluation experiments
- Tony Lee, Tuan Le: bwest tool automation
- Jiri Navratil, Ravi Prasad, Vinay Ribeiro: remote testbed users
- Grant Duvall, Nathaniel Mendoza, Brendan White: router config
- Kevin Walsh: CalNGI, NPRL access
  - Spirent SmartBits 6000 with SmartFlow software
  - Foundry Big Iron router
- Cisco: GSR12008 router
- Juniper: M20 router
- Endace: gigE DAG card for passive monitoring with NeTraMet and CoralReef
- Department of Energy SciDAC grant DE-FC02-01ER25466

Talk Outline

- Monitoring/Measurement goals
- Terms and Conditions
- Bandwidth estimation tools
- Evaluating and comparing tools
  - Lab tests with SmartBits
  - Lab tests with tcpreplay
- TeraGrid tests using the INCA architecture
- Future Directions

Why measure e2e available bandwidth?

- Configure overlay routes
- Select “best” content distribution server
- Adjust encoding rate on streaming applications
- Verify SLA and QoS
- Use as criterion for end-to-end admission control
- Construct a peer-to-peer application topology
- Select inter-domain egress ISP
- and…

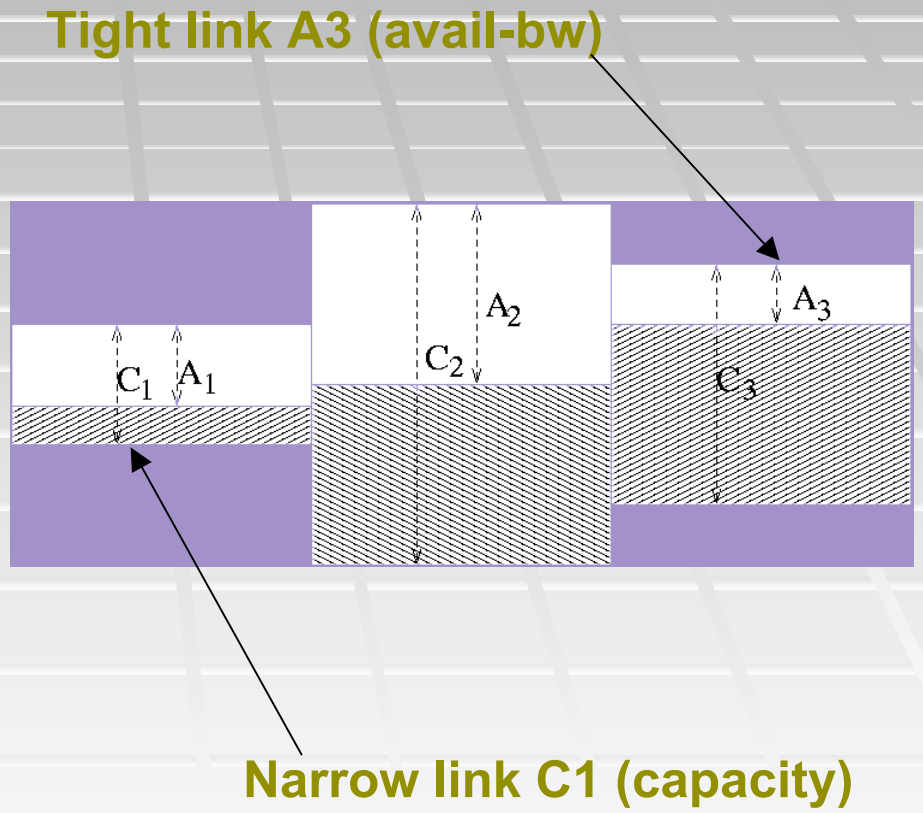
End-to-end performance perspectives

- User goals:
  - Optimize my application performance
  - Move my data… FAST
  - With whom am I sharing network bandwidth?
- Sysadmin goals:
  - Identify problems
  - Set realistic performance expectations
- Common denominator:
  - **Maximize** available bandwidth

Terms

"Bottleneck" is not a meaningful term

- e2e Capacity (C): min link capacity in the path
- e2e avail-bw (A): min unused bandwidth at time T
- BTC: max achievable TCP throughput



…and Conditions
(factors impacting e2e net performance)

- Cross-traffic (load level, burstiness)
- Traffic type (TCP/UDP) mix
  - We assume that 80%+ of apps are TCP
  - Number of competing streams
- Host TCP settings
- MTU size
- Clock synchronization
- Router buffer sizes and COS or QoS

Measuring end-to-end Available Bandwidth

- It's not easy, and tools haven't been validated.
  - Even fewer tools developed and validated on high speed links.

  CAIDA is performing first comprehensive tool evaluation on high speed links in CAIDA/SDSC lab.
- Well-known Iperf (persistent TCP connection w/ large advertised window)
  - Can be intrusive: can saturate the path and increase path delays and jitter…depending on time scale and if no limits on its bw use
  - Measures "brute force" avail-bw
- Pathload (Self-Loading Periodic Streams)
  - Attempts to be non-intrusive over time (uses < 10% avail-bw)
  - Measures the dynamics of avail-bw over time

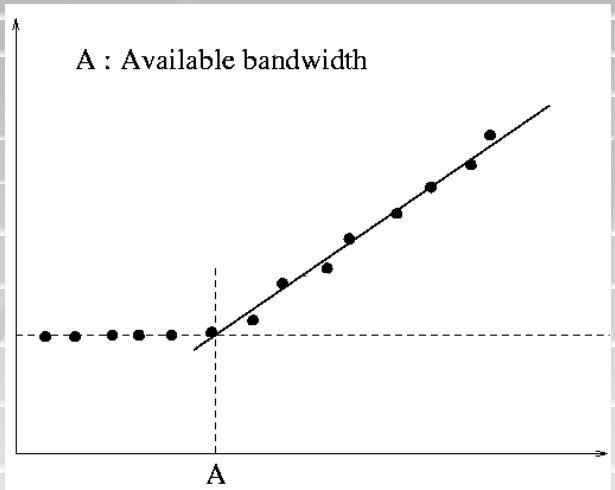
Current e2e Tools

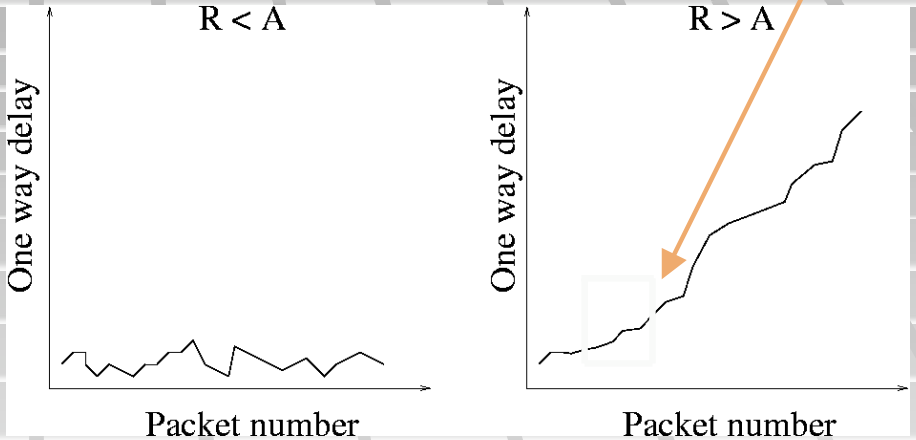
Tool Class	Tool	Authors	Methodology	Tool	Authors	Methodology
Per-hop Capacity	*clink* √	Downey	VPS	*pathchar* √	Jacobson	VPS
	pchar √	Mah	VPS			
End-to-End Capacity	*bprobe*	Carter	pkt pair	*pathrate* √	Dovrolis-Prasad	pkt pairs,train
	nettimer	Lai	pkt pairs	*sprobe* √	Saroiu	pkt pairs
End-to-End Available Bandwidth	*abing* √	Navratil	pkt pairs	*netest* √	Jin	unknown
	cprobe	Carter	pkt trains	*pathload* √	Jain-Dovrolis	SLoPS
	IGI √	Hu	SLoPs	*Spruce*	*Strauss*	*Mod. SLoPS*
Bulk Transfer Capacity	*cap*	Allman	emulate TCP tput			
	treno	Mathis	std TCP tput			
Achievable TCP Throughput	*iperf* √	NLANR	|| TCP connect	*ttcp*	Muuss	TCP connect
	Netperf	NLANR	|| TCP connect			

Sidebar: How pathload works…

…find the range of the knee

Concept:





- Send ≈100 probes of equal-sized packets at rate R and measure one-way delays; iterate while modifying R (and limit probing rate to < 10%)
- One-way delays only increase when the stream rate R is *larger* than the avail-bw A

Summer 2004 Tool Eval

- Tools to test
  - **Pathload**
  - **Pathchirp**
  - **Spruce**
  - **Abing**
  - **Iperf**
- Performance metrics
  - Error
  - Overhead traffic
  - Time to measure
- Testing metrics
  - Test frequency
  - Test scheduling

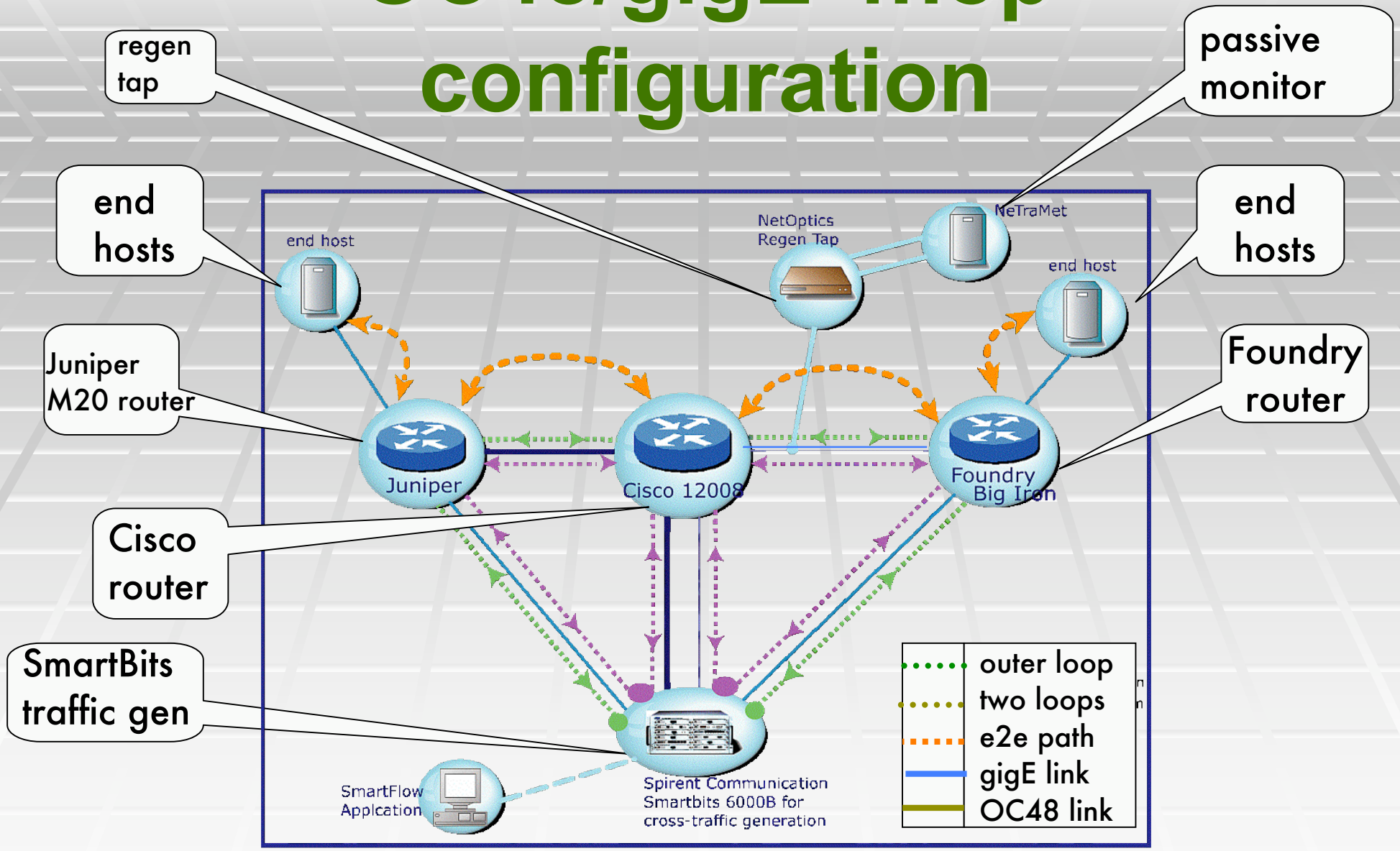
- Tools not (or no longer) under test
  - Abw
  - Bprobe
  - Cprobe
  - Clink
  - Pathchar
  - Pchar
  - Pipechar
  - tracerate

Lab Tests with SmartBits

- Use reproducible test conditions
- Can test against saturated links
- Validate tool and cross traffic
- Test “black box” e2e tools against same scenarios
  - Identify conditions where tools work well
  - Give developers an environment for refining their tools

(synthetic traffic, and unresponsive to TCP)

OC48/gigE 4hop configuration



Lab Tests with tcpreplay

- Use the same anonymized trace for all tools
- Estimate the load level using CoralReef
- One end host generates tcpreplay cross-traffic
- Separate end host runs the tool under test

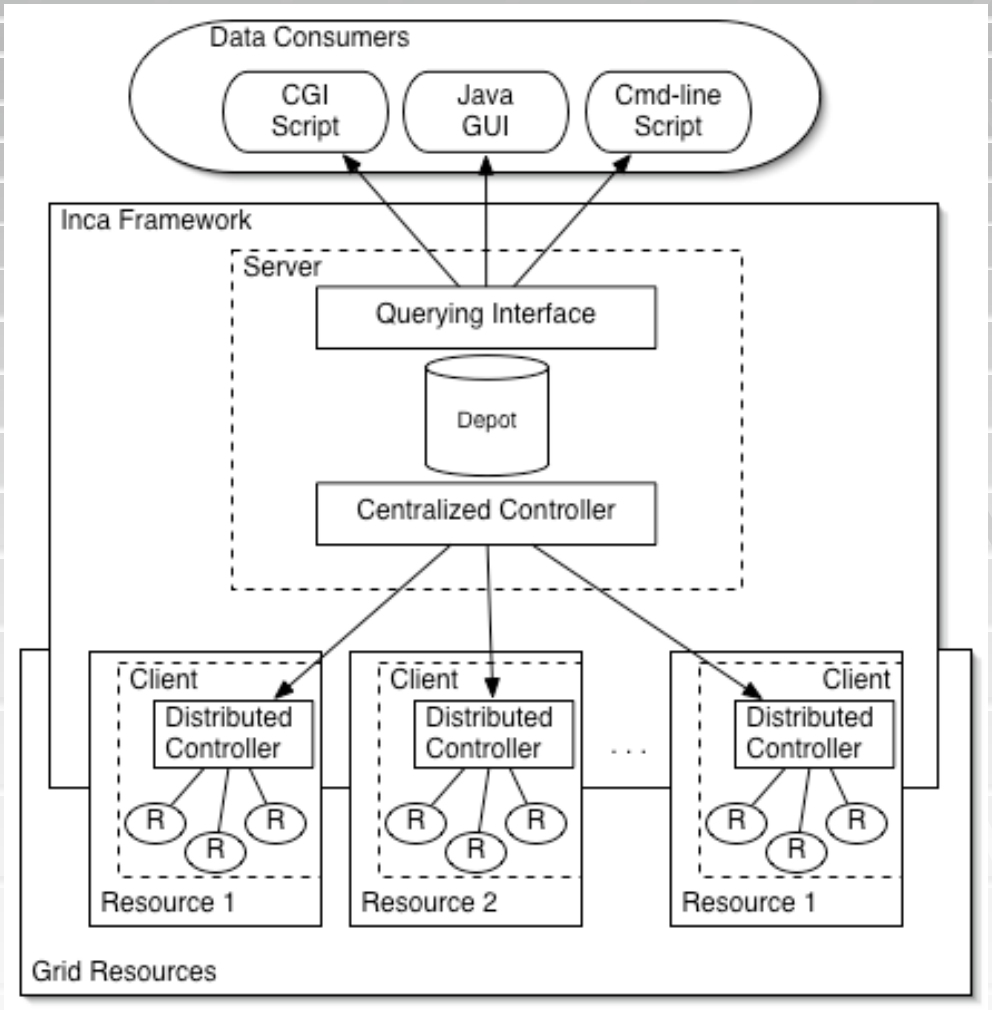
  (real traffic, but unresponsive to TCP)

Tests with real traffic

- INCA Test Harness and Monitoring Infrastructure http://tech.teragrid.org/inca/
- Take advantage of INCA's:
  - Full mesh deployment
  - Data repository/archive
  - Web interface
  - Scheduling options
- To collect network performance data:
  - Add Network Reporter
    - Reporter-Pair - a new variation
    - Same wrapper can work with multiple avail-bw tools

Inca Architecture

- Data consumer - user-friendly web interface, application, etc.
- Framework - daemons
  - Planning and execution of reporters
  - Centralized data collection
  - Publishing
- Reporter - a script or executable

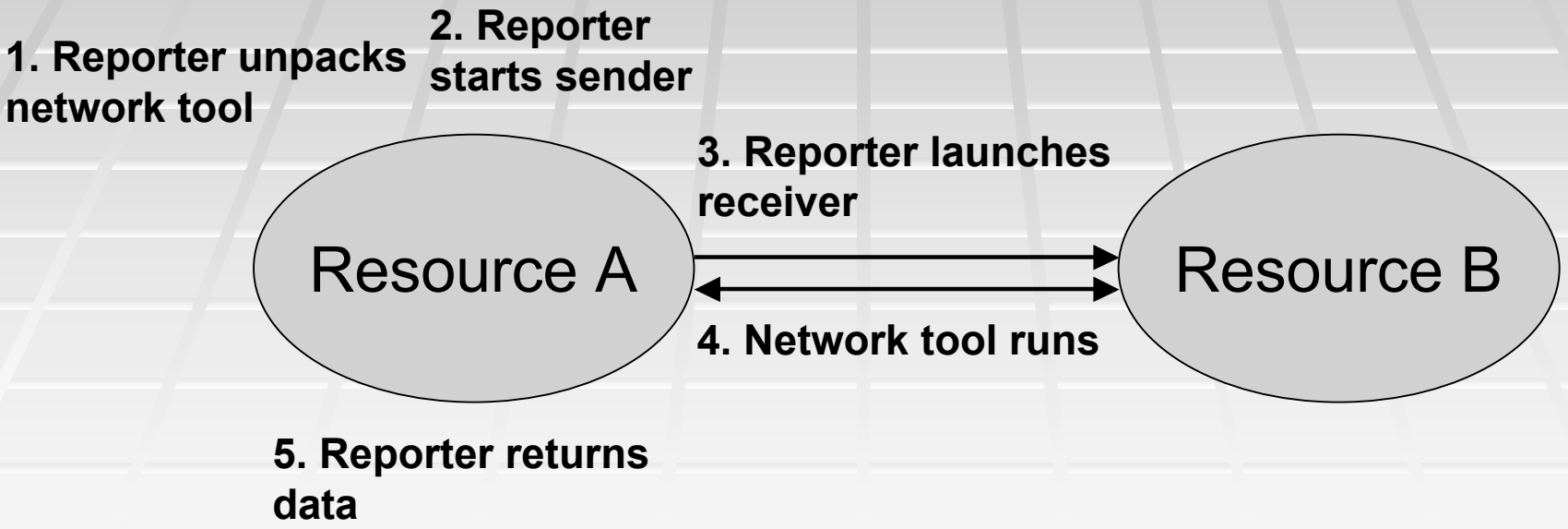


Gathering performance data

1. Write reporter to wrap benchmark and print XML output according to Inca reporter specification
2. Write configuration file to express:
   a) Inputs
   b) Frequency of execution
   c) Data to archive
3. Write web page to display data

Writing performance reporter

- Perl API to enable running of network probes across sites (uses globusrun)

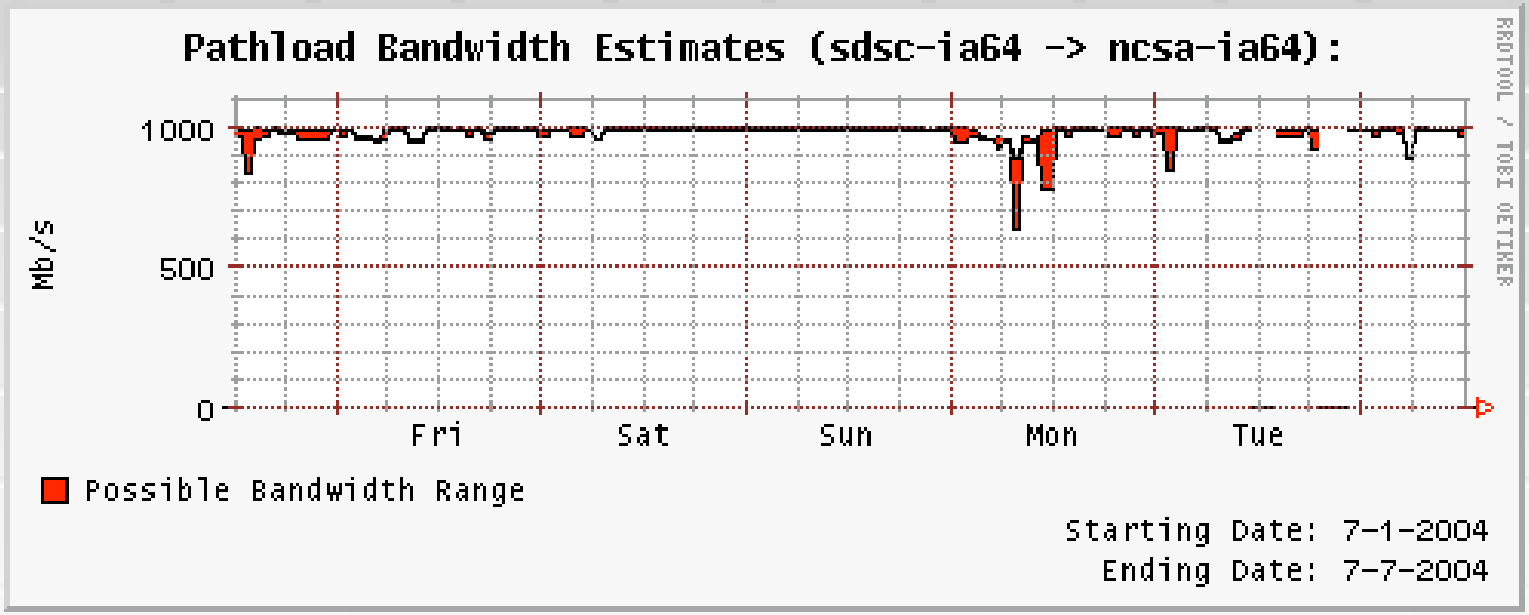


Executing reporter

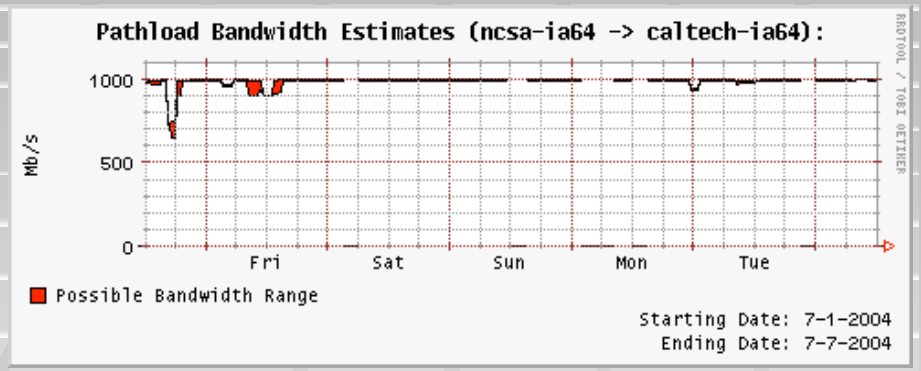
- Now: Cron scheduling
  - Schedule far enough apart so they don't collide
  - Not foolproof
- Move to token-passing protocol (NWS)?

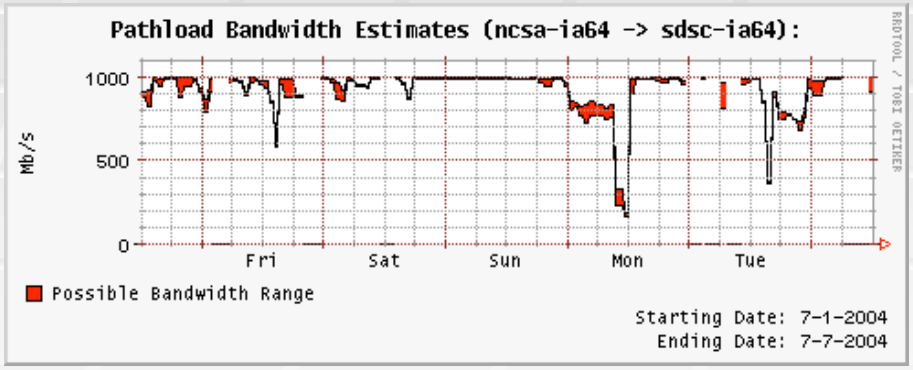
Graphing data

- Calls rrdtool commands to generate graphs
- CGI script currently uses SOAP call to get graph from Inca archive



More graphs from CGI form





- User selects:
  - Source
  - Dest
  - Start date/time
  - End date/time
- Planned:
  - Weather map style

Future Directions…

- Justify test scheduling frequency
  - Now: once/hr
  - Check result distributions
  - Refine scheduling: Move to token-passing protocol (NWS)?
- Compare results of multiple tools
  - pathload, pathchirp, Spruce, iperf
  - Consider error and overhead
- Refine graphs and web interface
- Run network probes across different OSes
- Consider more e2e paths than just between login nodes

(especially aggregated bandwidth between site gridftp servers?)

Discussion: SOBAS for apps

- Socket Buffer Auto-Sizing (SOBAS) [Prasad, Jain & Dovrolis, GaTech]
  - Apps use a SOBAS enabled socket library.
  - Concept: Limit the send window after reaching avail-bw to avoid “self-induced” packet loss.
  - Experimental results show 20-80% increase in throughput compared to TCP transfers using max possible socket buffer size.

R. Prasad, M. Jain and C. Dovrolis, “Socket Buffer Auto-Sizing for High-Performance Data Transfers” Journal of Grid Computing June 2004. http://www.cc.gatech.edu/~ravi/tools/sobas.tar.gz

Summary

- CAIDA is evaluating bwest tools in both lab and real high-speed environments.
- TeraGrid's INCA architecture now supports available bandwidth measurements.
- Pathload reports a range variation of available bandwidth on an e2e path.
- INCA/pathload measures available bandwidth on TGrid e2e paths (login node to login node).