Misure di prestazione della rete: l'esperienza INFN

Alcune considerazioni teoriche e pratiche sui numeri della rete e su come misurarli

Luca Carbone
INFN Milano (a Bicocca)

I numeri della rete

- **Bande** (nel senso **informatico** del termine)
  - **Capacita'** di un link o di un path (=successione di link)
  - **Traffico** su un link (o su un path?)
  - Banda residua su un link o un path
  - **Throughput massimo** su un link o un path
- **Tempi**
  - **Latenze** (RTT, one way delay)
  - **Disponibilita'**
- **Numeri vari...**
  - **Pacchetti**: persi, corrotti, dimensione, ...

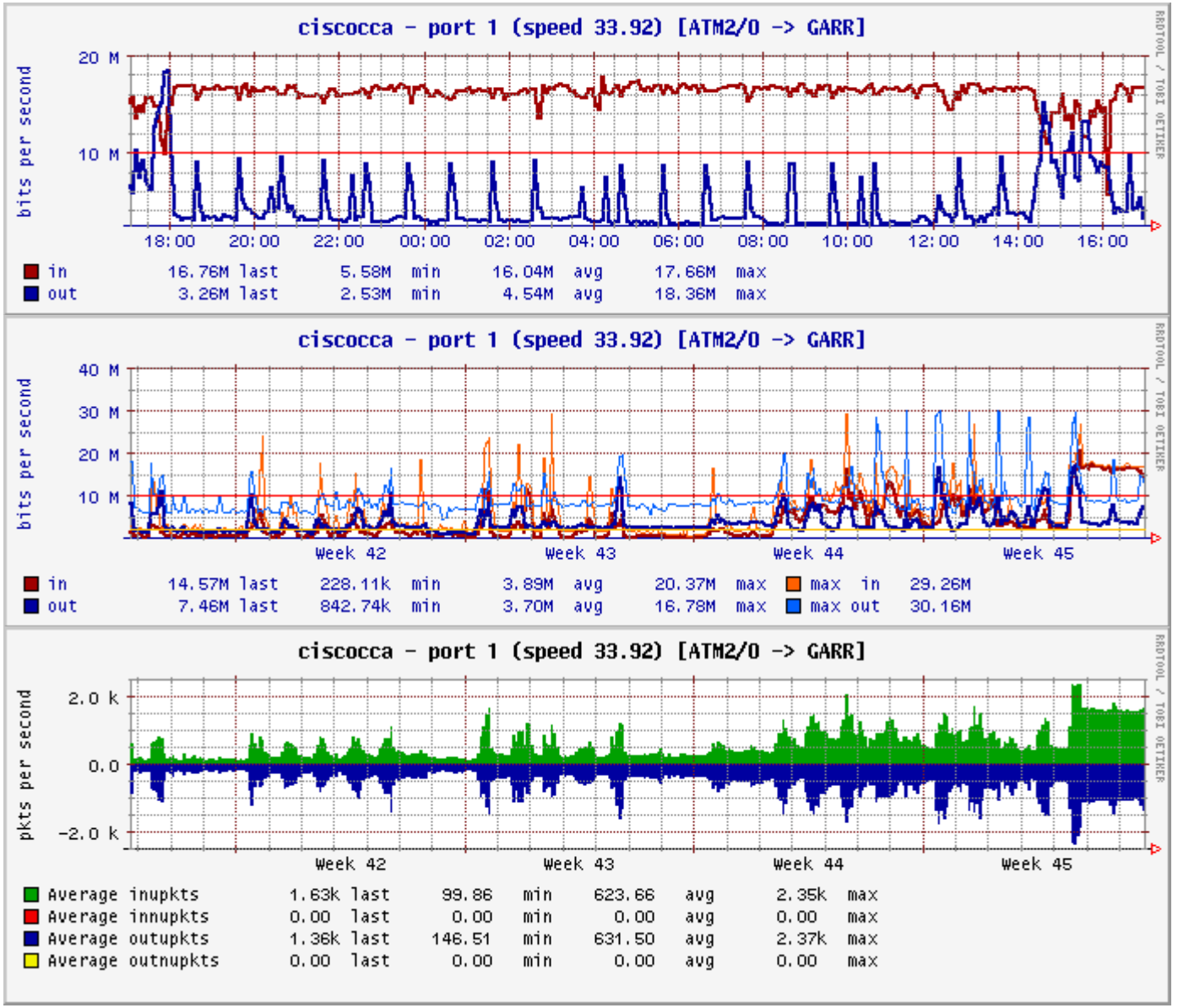
La 'banda'

- Capacita' e traffico sono concetti relativamente ben definiti (quasi sempre: BEA, BGA, overhead di protocollo, ...), noti o misurabili senza grande sforzo. La capacita' complessiva di un path, per esempio, non potra' essere maggiore di quella del suo link piu' lento; il traffico su un path e' gia' piu' sfuggente.
- Banda residua: la differenza tra la capacita' di un link ed il traffico su esso presente. Nel caso di un path completo, la minima tra le bande residue dei link che lo compongono. Ma come e' correlata al throughput utente?

Lo standard **de facto**

- Traffico & pacchetti su un link: SNMP
- **Misure passive** (MP):
  - SNMP agent sugli apparecchi di rete
  - data collector (DB storico) su macchina dedicata:
    - MRTG + rrdtool
    - Cricket + rrdtool
- Semplice, flessibile, estensibile (ottetti, pacchetti, errori; poi carico CPU, temperature etc.), poco/non intrusivo, 'abbastanza' efficiente (~migliaia di bersagli/minuto)

Qualche esempio ...



MP SNMP: le limitazioni

- Accesso limitato ai solo nodi locali.
- Anche avendo accesso ai nodi remoti, non ho nessuna indicazione sul comportamento end to end:
  - Se il link locale e' scarico, perche' il transfer rate verso il server X e' cosi' basso?
  - Posso scaricare i dati da N server differenti: quale scelgo?
  - Come ottimizzo le mie applicazioni (in senso lato)?
- Il mio provider rispetta i termini del contratto?

MP – un passo in piu': sniffer et al.

- tcpdump, ethereal, ntop (o NetFlow) + Data Collector (il solito rrdtool) + analisi ad hoc
  - opportunamente disposti sulla rete possono collezionare dati importanti: flussi (numero e tipo: src, dst, porte, ...), dimensione pacchetti, tempi di arrivo, ...
  - l'analisi (offline) dei dati permette di ricavare informazioni su traffico, prestazioni end to end e non solo (banda delle linee attraversate, congestione, stabilita' nel tempo, top talkers etc.)

Inter-arrival times distribution: un esempio (quasi) reale...

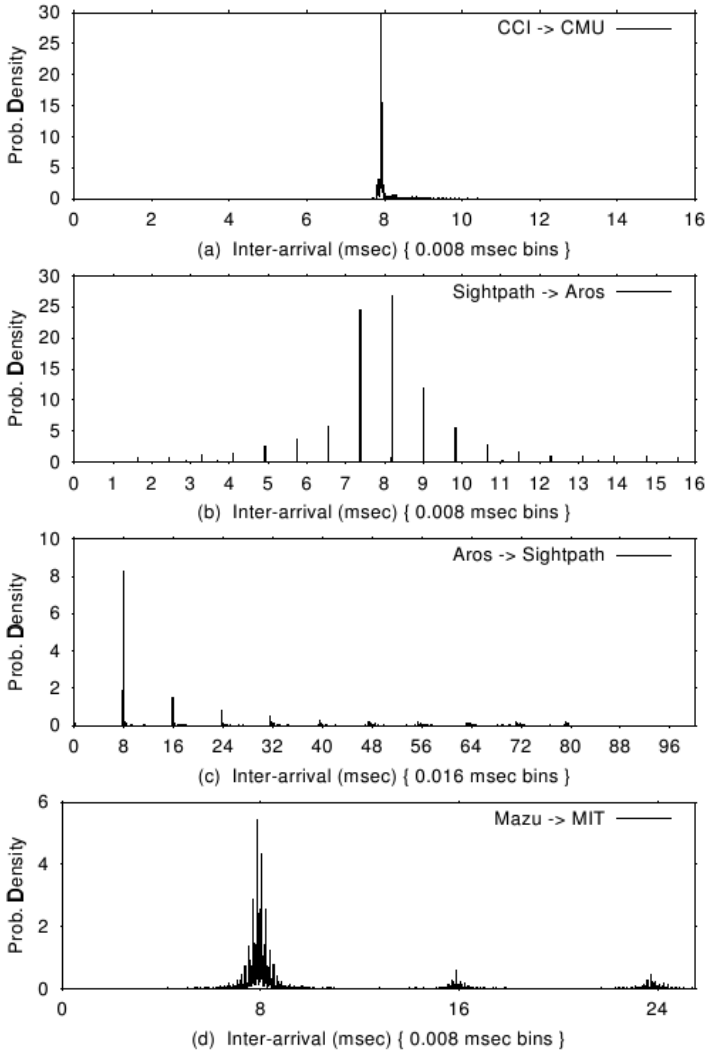
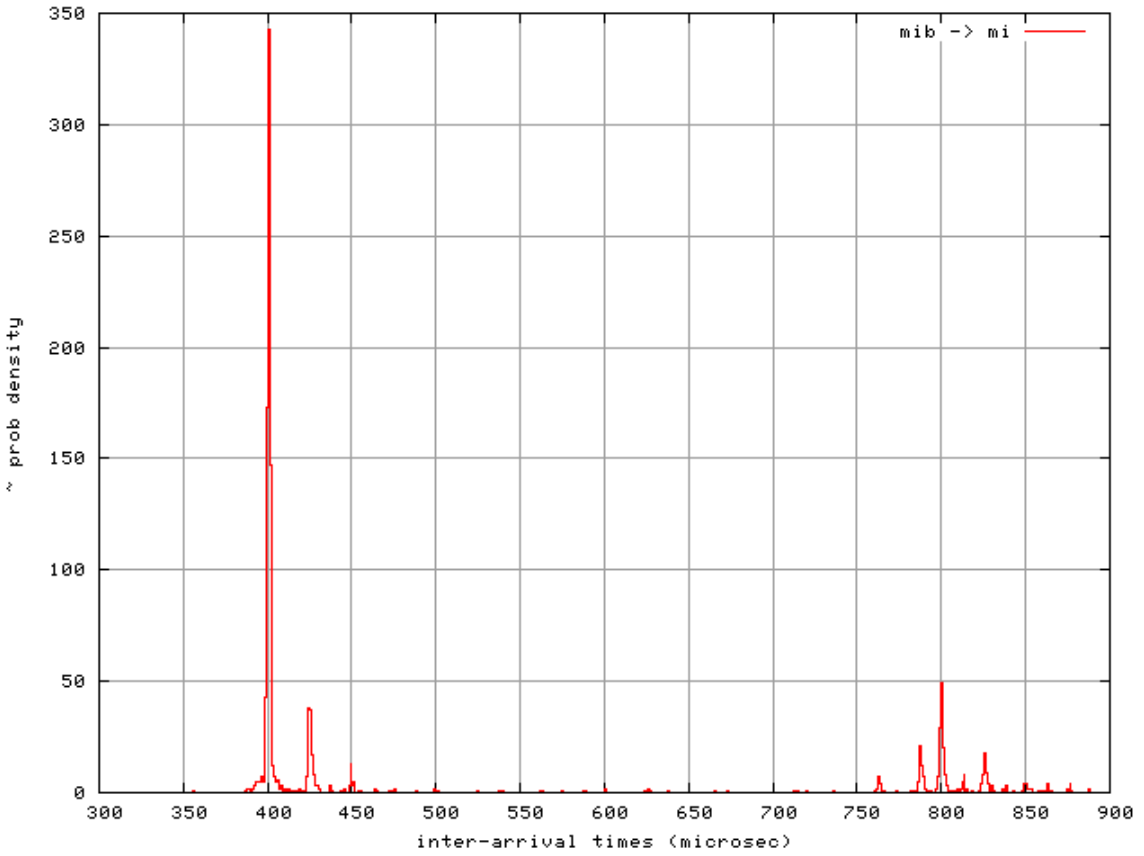
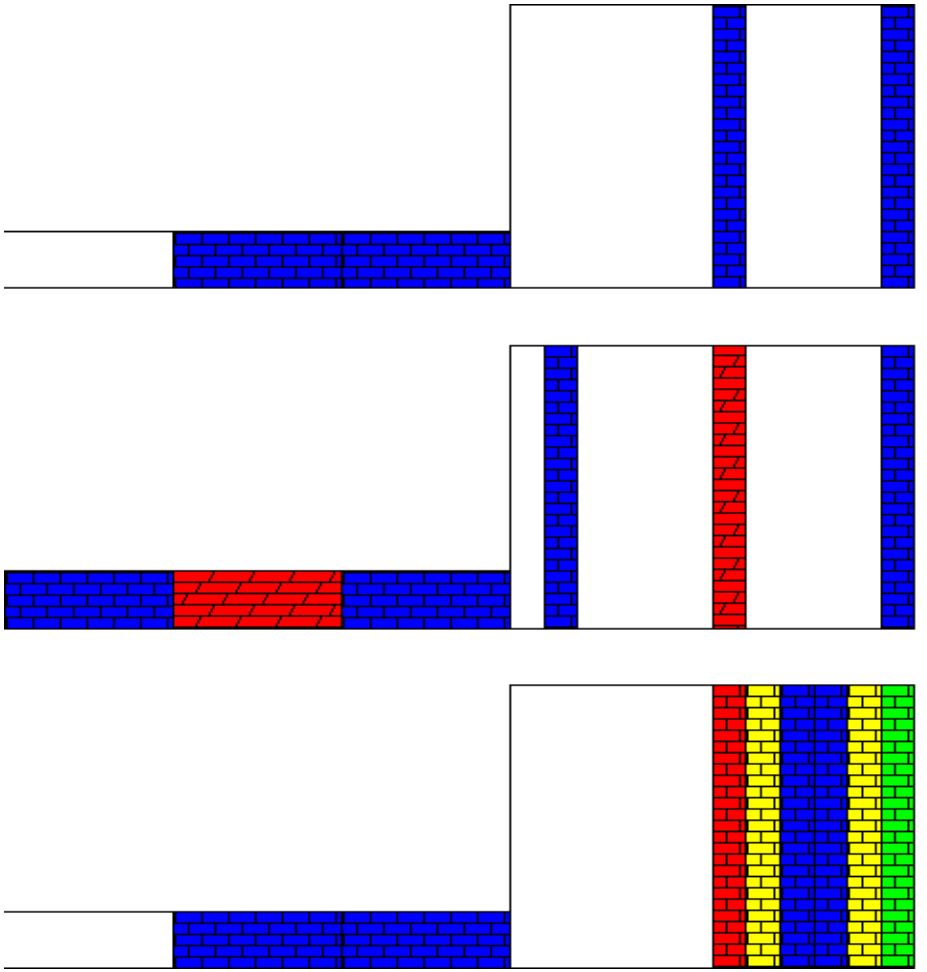


Figure 1: Common patterns in the PDF of inter-arrival times in a single TCP flow. (a): a single spike; (b): a spike bump; (c): a spike train; (d): a train of spike bumps.



... e la sua interpretazione

- Successione di due link di capacita' differente: 34 Mbps -> 1Gbps
- Le strutture principali indicano accodamento (possibile congestione) sul link lento; i picchi piu' piccoli intorno ai principali accodamento sul link veloce.
- La situazione inversa (link veloci -> link lento) cancellerebbe di fatto alcune strutture (modo dominante = ultimo link)



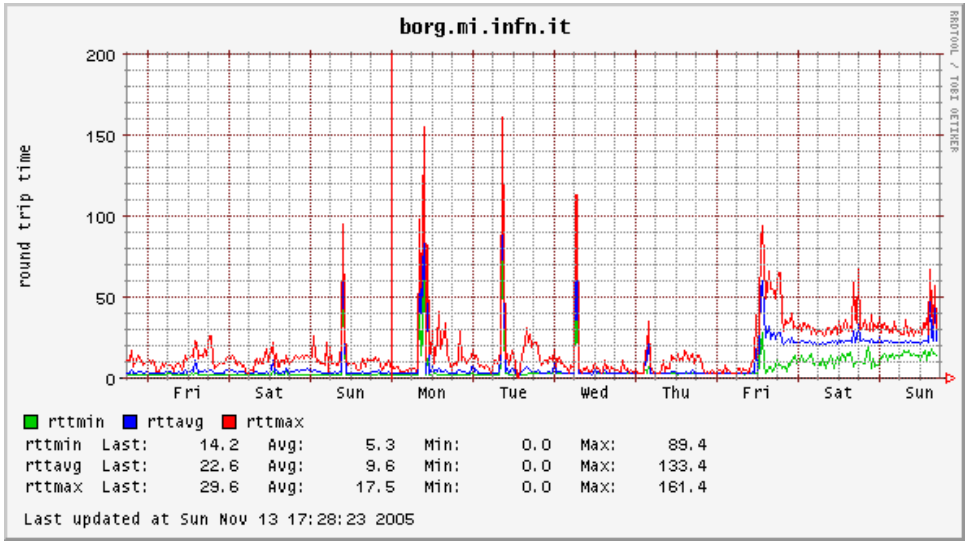
MP: SNMP polling + sniffer

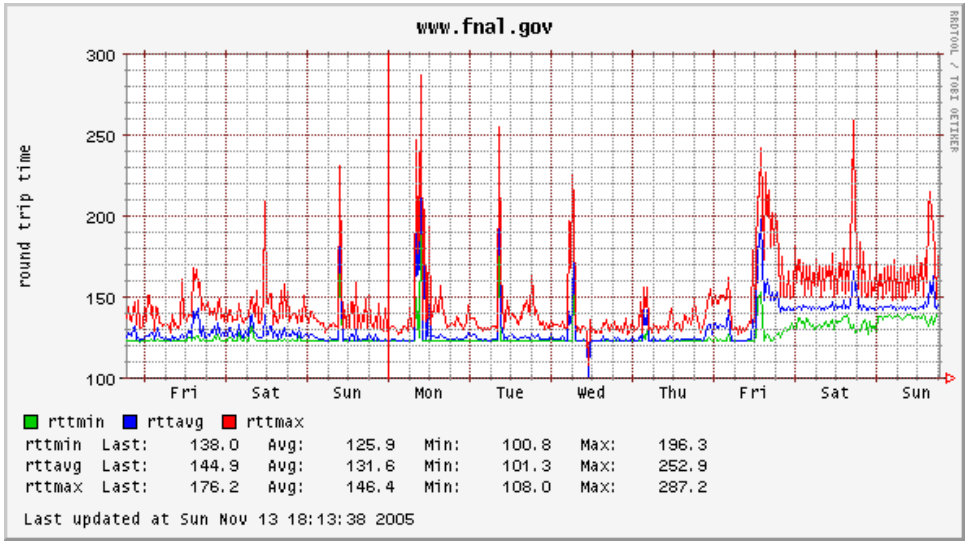
- Impatto nullo (o quasi) sulla rete
- Fotografia accurata della situazione locale
- Descrizione dettagliata del traffico in entrata ed uscita
- Analisi a posteriori (offline) non sempre (quasi mai?) semplice (NO reverse engineering dei router...); non mirata; sempre piu' pesante al crescere del traffico osservato e del dettaglio richiesto.

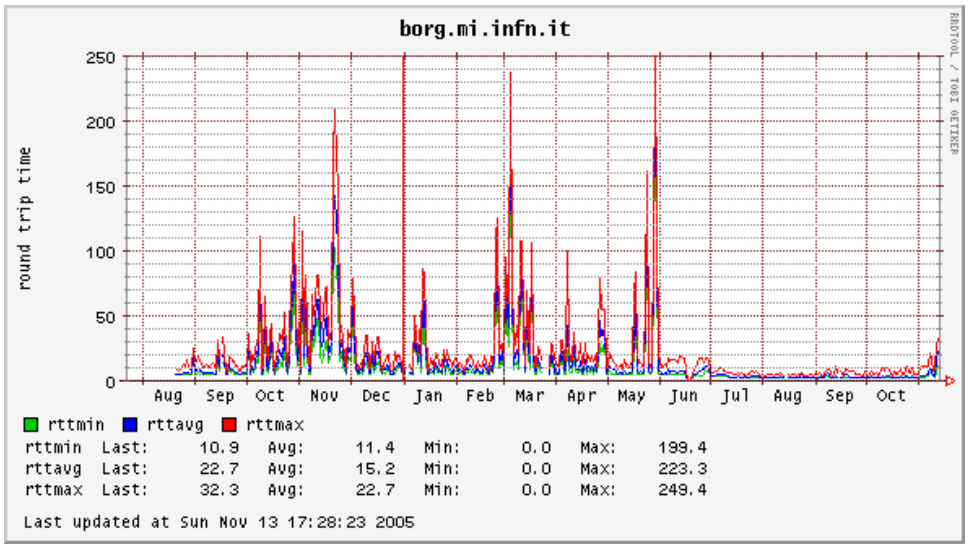
Misure attive (MA)

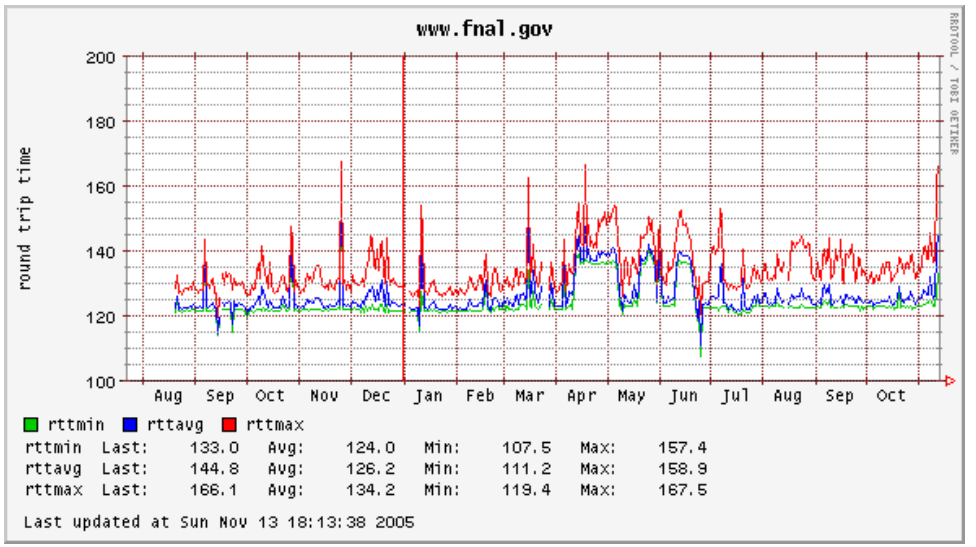
- Si inietta sulla rete traffico di caratteristiche note per studiarne il comportamento:
  - ICMP probes (ping) per il rilevamento di rtt, one way delay (OWD), disponibilita' nel tempo;
  - Flussi TCP (singoli, paralleli, mem2mem, disk2disk – ttcp, iperf, netperf, ...) per la misura del throughput;
- I treni di ping probes periodici abbinati ad un data collector sono l'opzione piu' semplice e di piu' facile implementazione per iniziare ad abbandonare la visione 'locale' delle MP.

ping + cricket

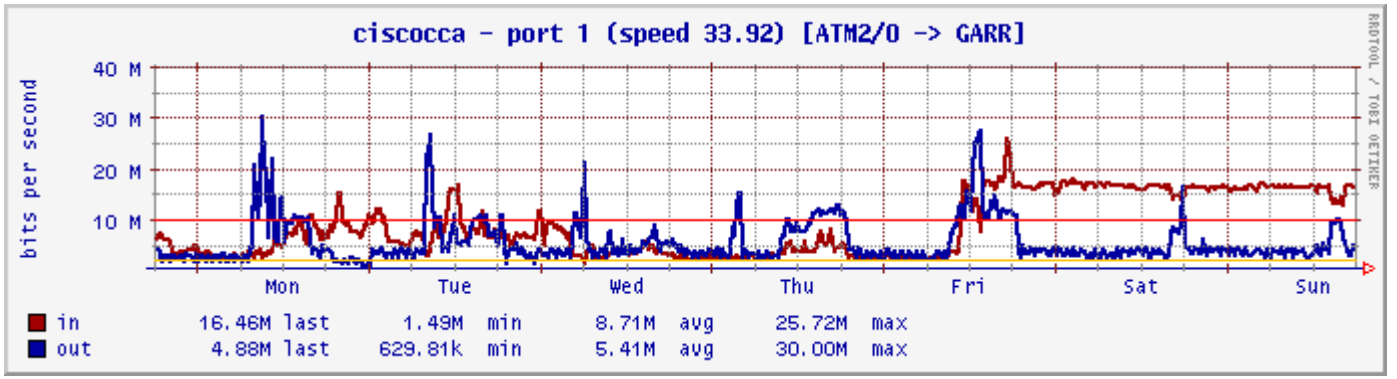


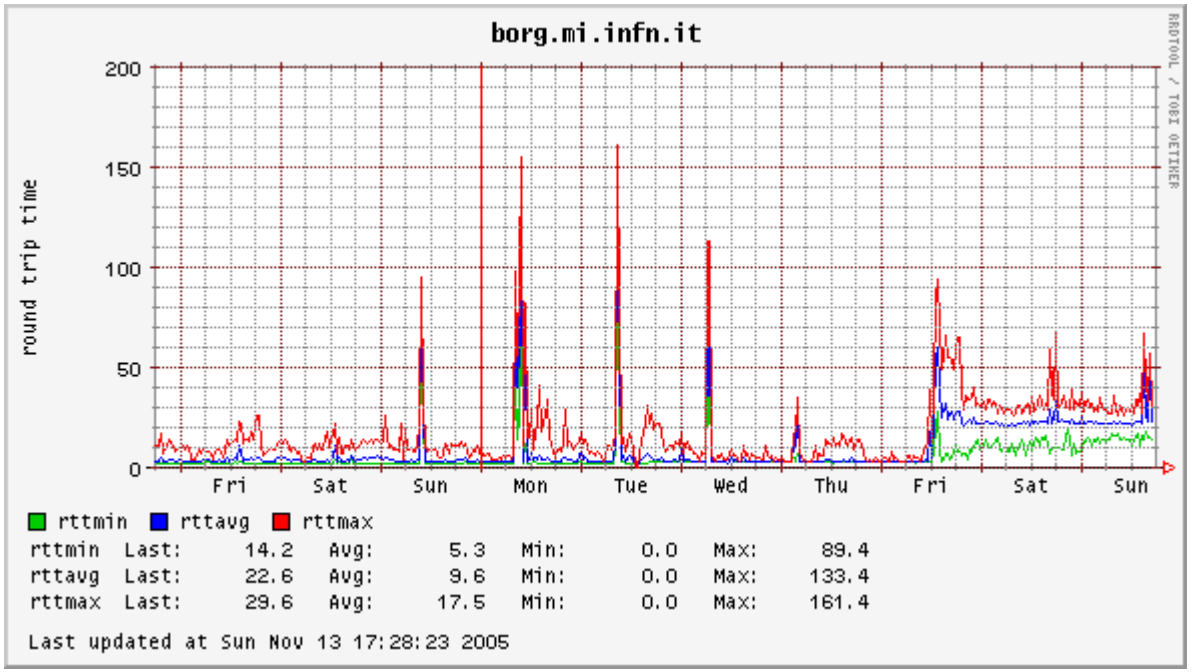




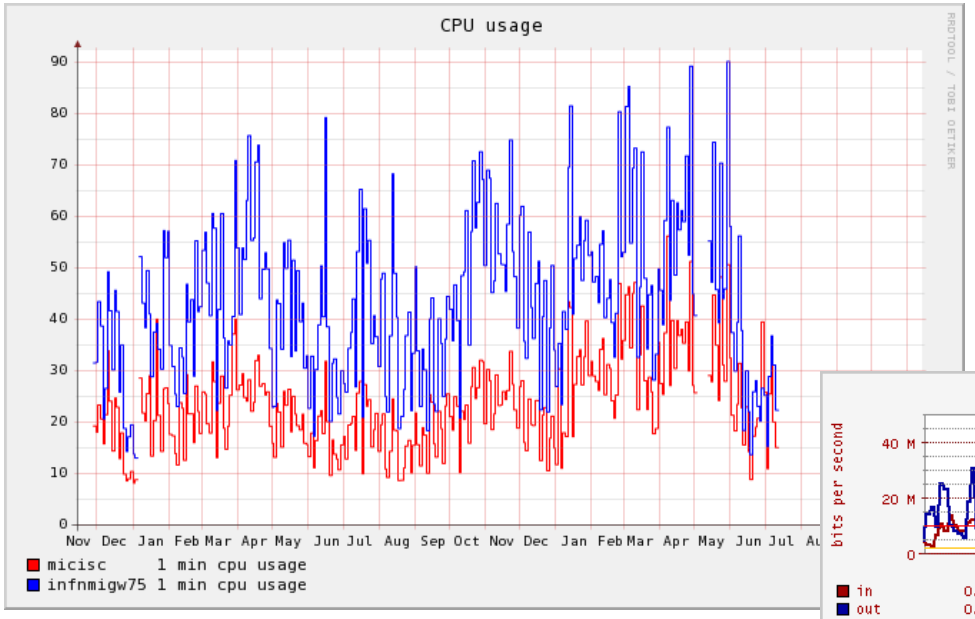


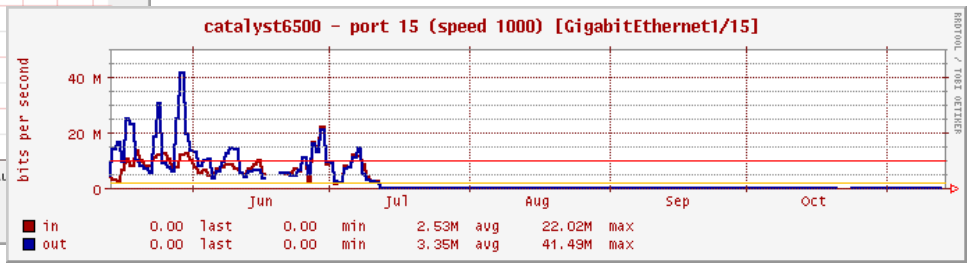
Correlazioni ping/SNMP (1):

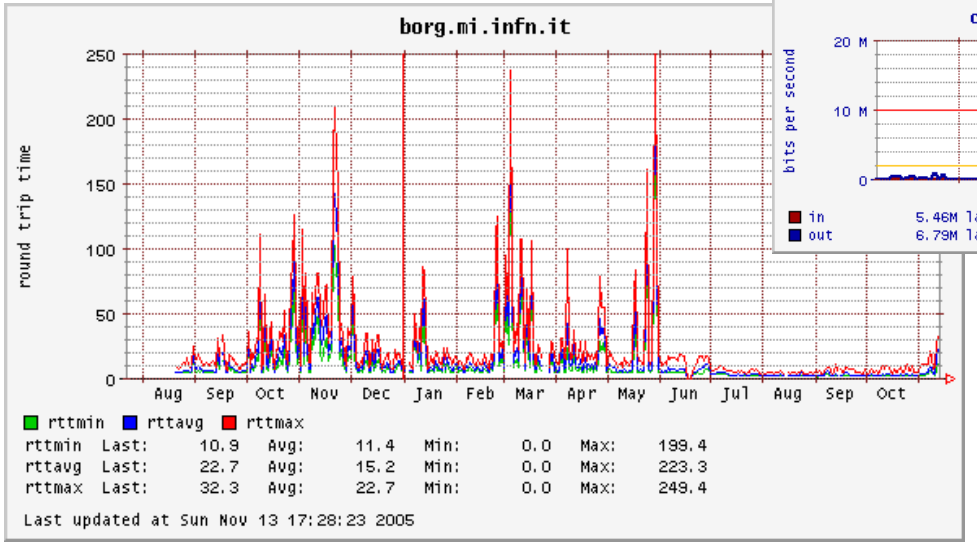


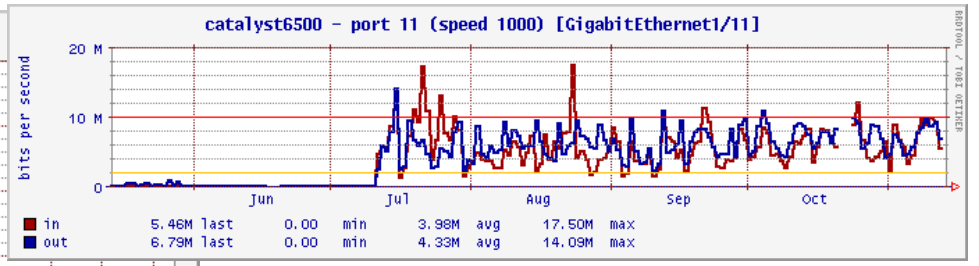


Correlazioni ping/SNMP (2):









E' sufficiente?

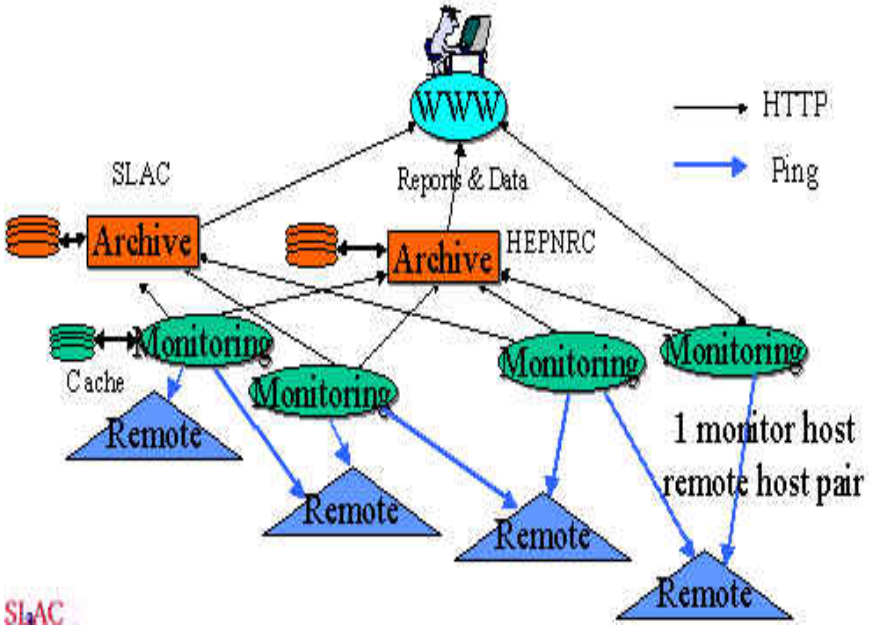
- Comodo e semplice ed efficiente; funziona anche in presenza di sistemisti scarsamente collaborativi nei siti da controllare; fornisce risultati qualitativamente, ma non quantitativamente, correlati ad altre metriche interessanti (traffico, tortuosita' di un path), ed a volte ambigui (la congestione e' sul percorso di andata o di ritorno?)
- Il suo utilizzo su macchine di produzione e' possibile (il kernel si occupa direttamente dell'ICMP echo) ma puo' aumentare l'ambiguita' dei risultati

Soluzione: una rete di misura per la rete

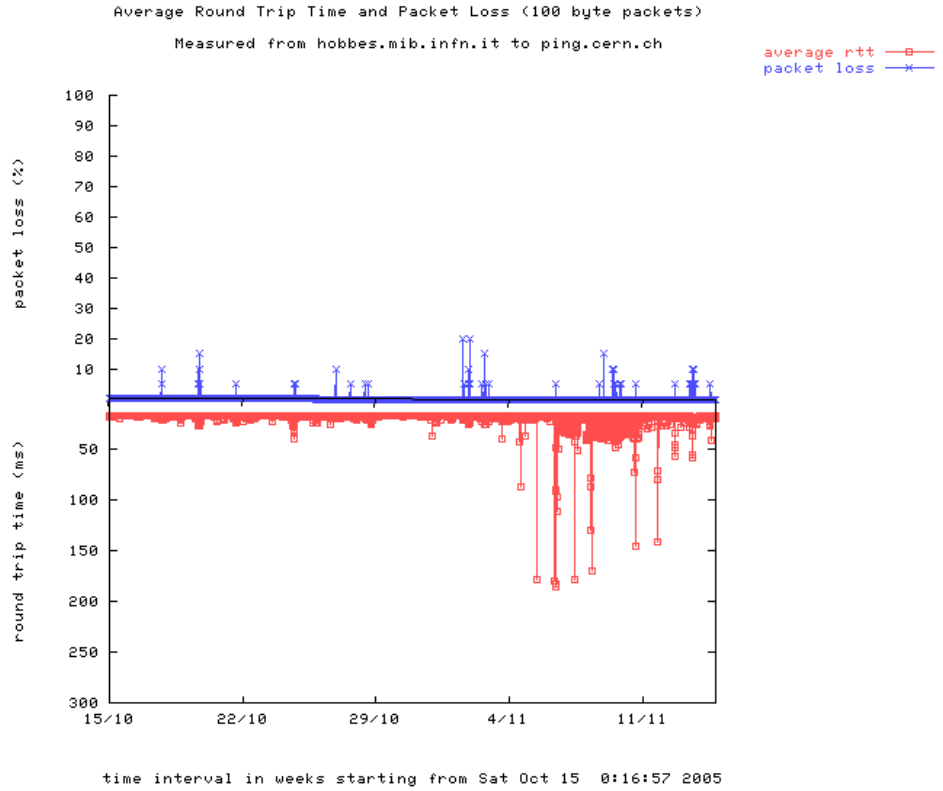
- Sembra opportuno creare un'infrastruttura ad hoc esclusivamente dedicata alla misura delle prestazioni reali della rete (punto di vista piu' interessante per l'utente finale e lo sviluppatore di applicazioni) ed alla verifica del suo funzionamento (punto di vista del network manager a qualsiasi livello)
- disponendo di un'infrastruttura dedicata si puo' anche estendere il numero delle metriche rilevate mediante tool standard o meno (cioe' da sviluppare, sperimentare, validare).

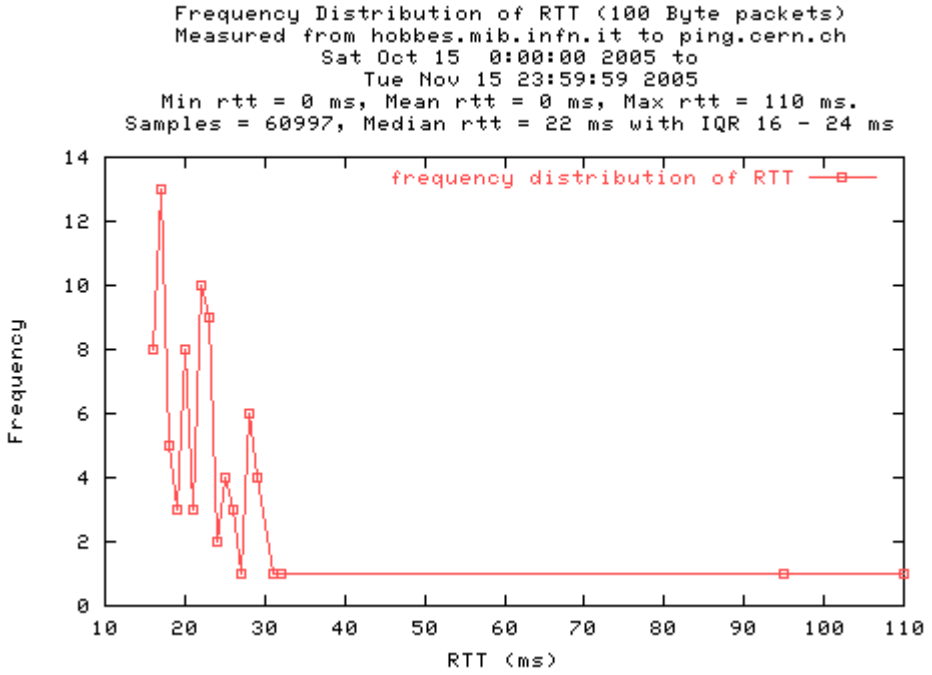
The PingER project

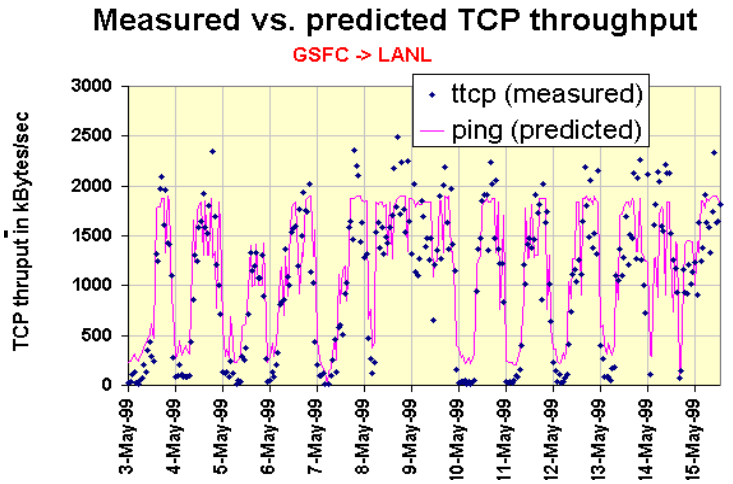
- Attivo dal 1995, e' basato su una architettura gerarchica composta da monitoring sites e remote sites (PC standard, piu' di 700 in 123 nazioni).
- Parametri misurati: RTT (due serie di 21 pacchetti da 100/1000B ogni 30 minuti: basso impatto), packet loss. Noti questi, tramite la formula di Mathis (e molta fiducia) si puo' stimare il throughput TCP



PingER data

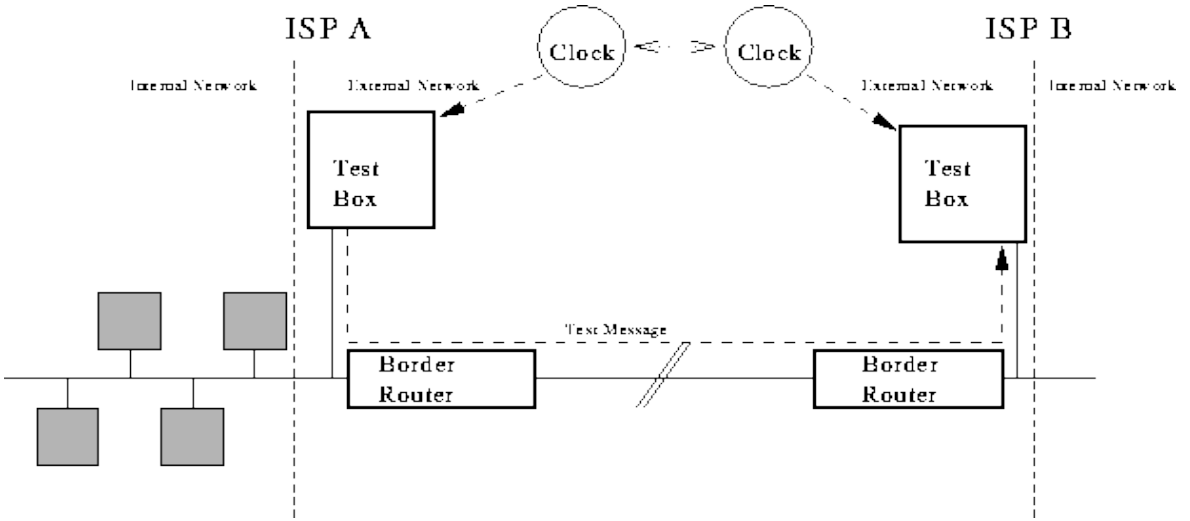

Average Round Trip Time and Packet Loss (100 byte packets)
Measured from hobbes.mib.infn.it to ping.cern.ch
average rtt
packet loss
packet loss (%)
round trip time (ms)
100
90
80
70
60
50
40
30
20
10
50
100
150
200
250
300
15/10
22/10
29/10
4/11
11/11
time interval in weeks starting from Sat Oct 15 0:16:57 2005


Frequency Distribution of RTT (100 Byte packets)
Measured from hobbes.mib.infn.it to ping.cern.ch
Sat Oct 15 0:00:00 2005 to
Tue Nov 15 23:59:59 2005
Min rtt = 0 ms, Mean rtt = 0 ms, Max rtt = 110 ms.
Samples = 60997, Median rtt = 22 ms with IQR 16 - 24 ms
frequency distribution of RTT
Frequency
0
2
4
6
8
10
12
14
10
20
30
40
50
60
70
80
90
100
110
RTT (ms)

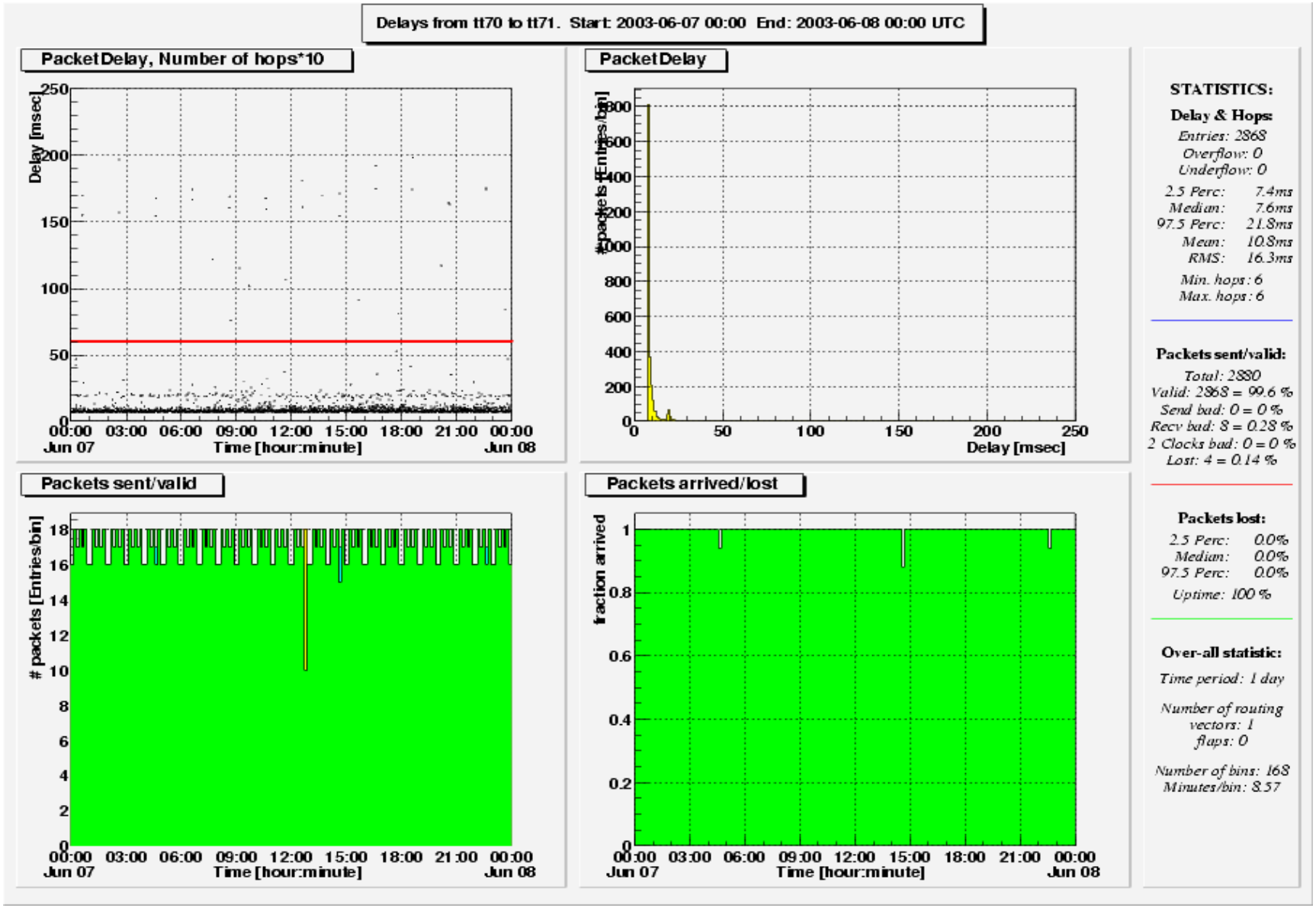

Measured vs. predicted TCP throughput
GSFC -> LANL
ttcp (measured)
ping (predicted)
TCP thruput in kBytes/sec
3000
2500
2000
1500
1000
500
0
3-May-99
4-May-99
5-May-99
6-May-99
7-May-99
8-May-99
9-May-99
10-May-99
11-May-99
12-May-99
13-May-99
14-May-99
15-May-99

RIPE-TTM

- Ogni nodo dedicato basato su H/W speciale (~2.5 Keuro, OpenBSD chiuso a chiave con RX satellitare per sincronizzazione clock) effettua misurazioni (one way delay, packet loss) verso tutti gli altri. Basso impatto sulla rete, specificatamente rivolto a ISP. Se ne prevedeva una discreta diffusione, non ha avuto molto successo.



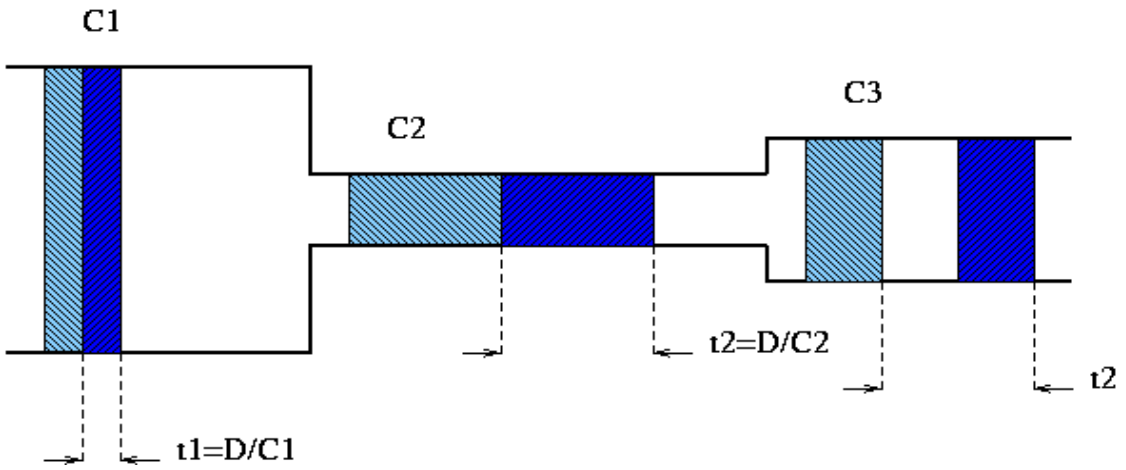
RIPE-TTM data

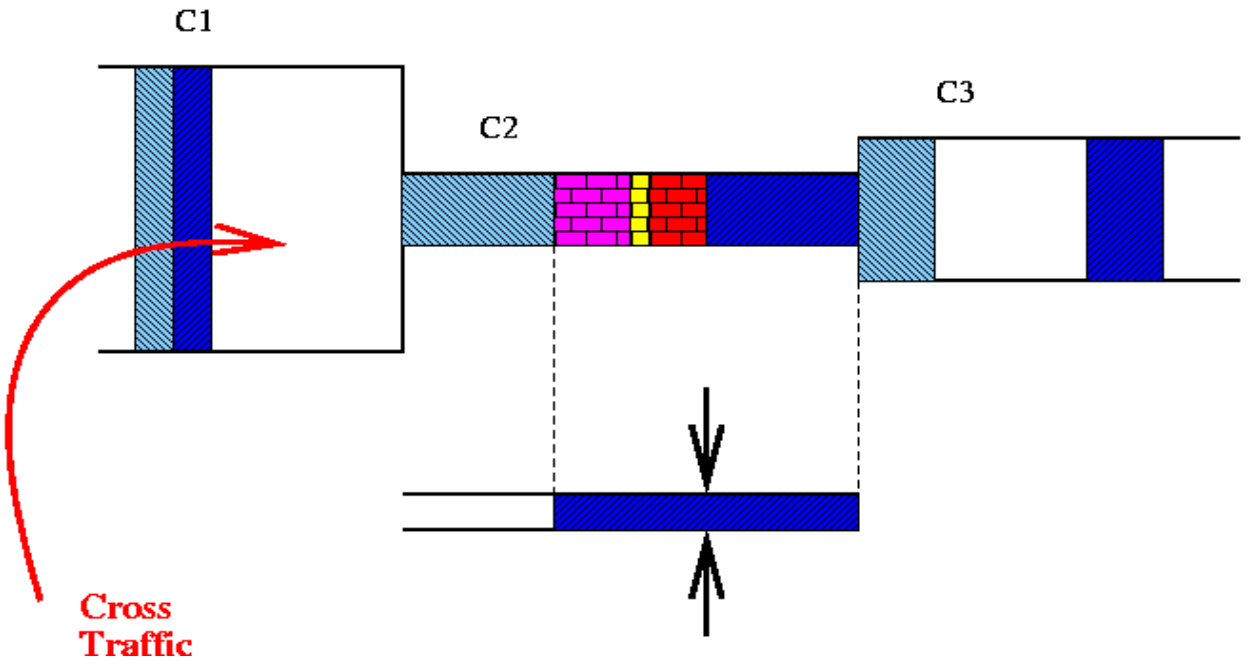


MA oltre il ping

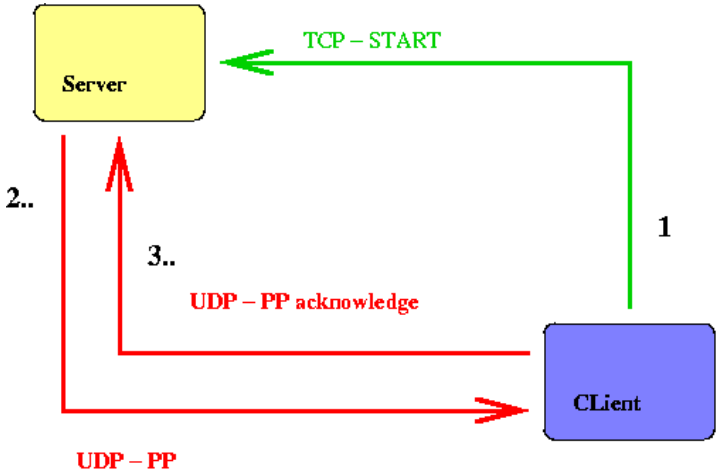
- Sembra utile integrare i dati di RTT con misure di banda. Ma tali misure:
  - possono essere intrusive: sottrarre banda a chi la stia lecitamente usando per dirgli quanta ne potrebbe usare puo' essere male interpretato...
  - possono essere tanto piu' intrusive quanto maggiore deve essere la granularita' della misura
- Problema: si puo' misurare la banda senza consumarla?

Packet Pair/Train Dispersion



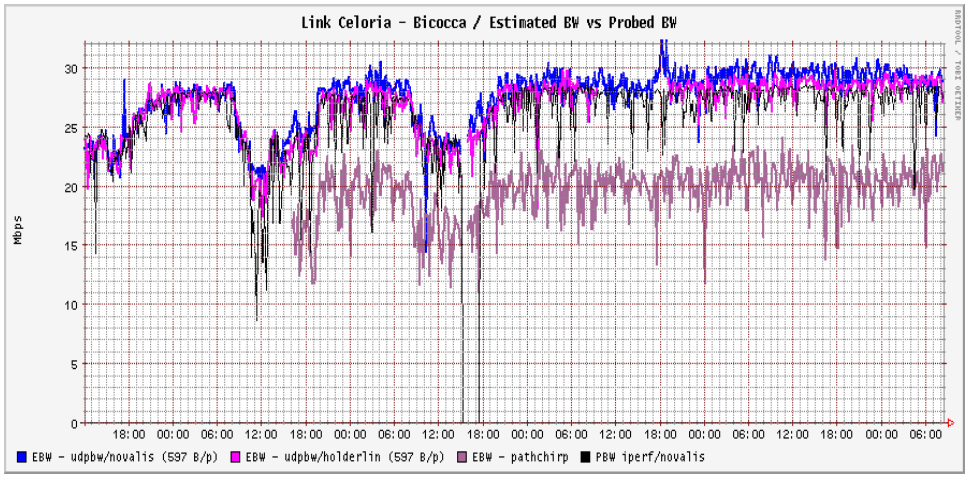


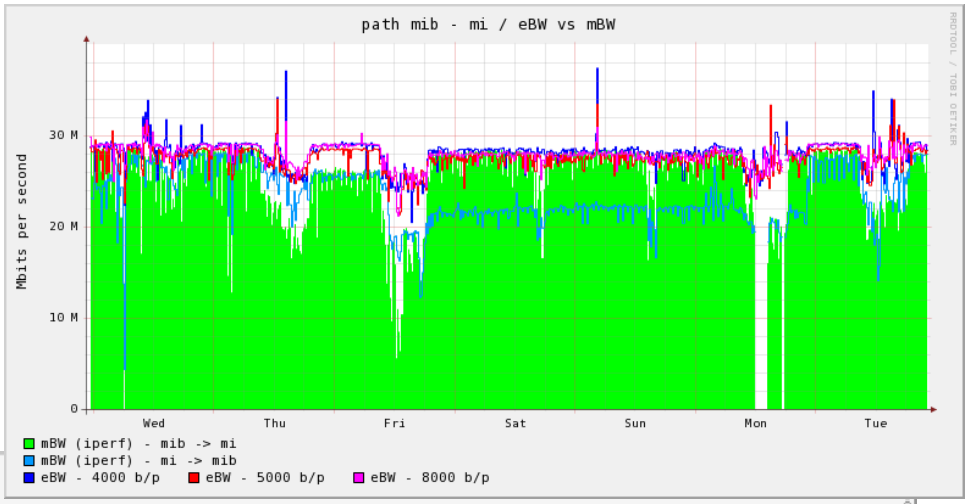
PPST (PP Simple test)

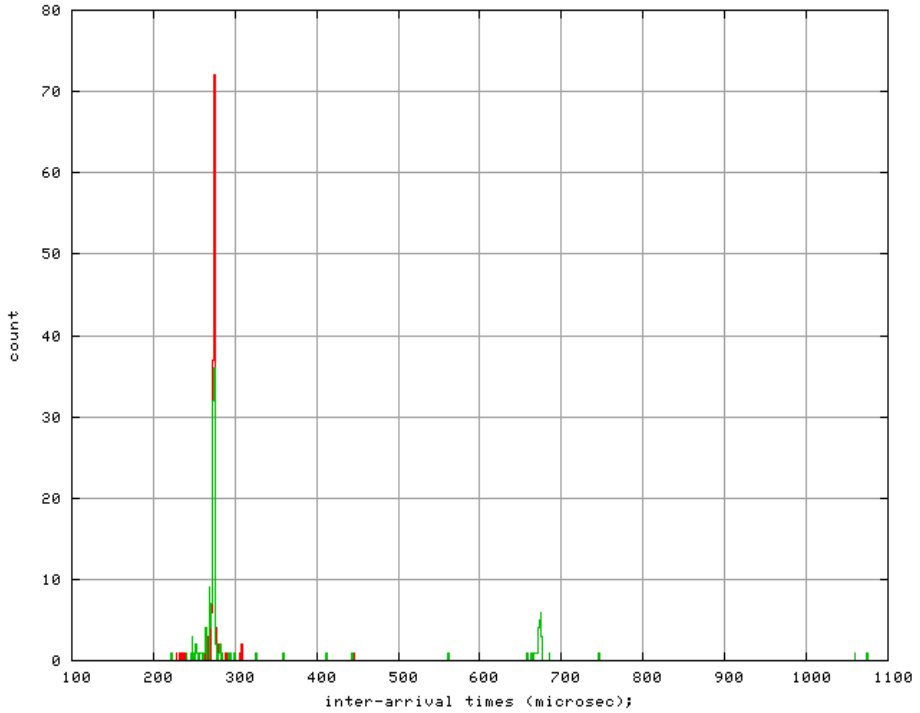


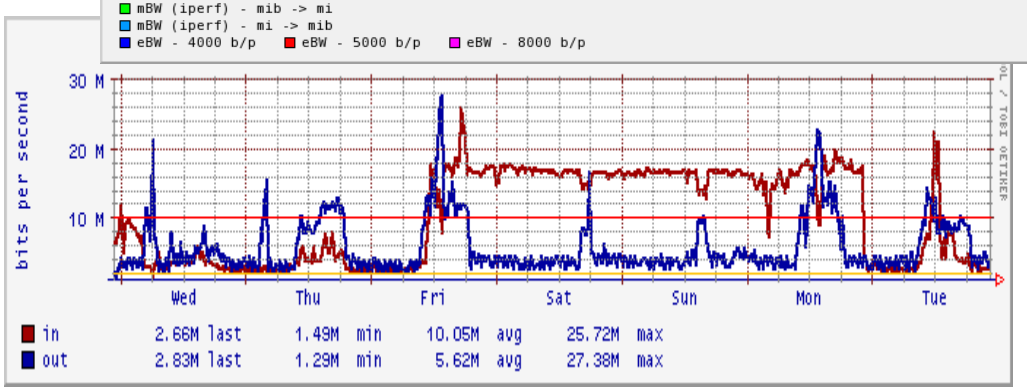
Parametri: dimensione pacchetti, numero coppie, porte, etc. etc.

Output: ritardi singoli, ritardo medio, numero coppie ricevute, durata totale del test.









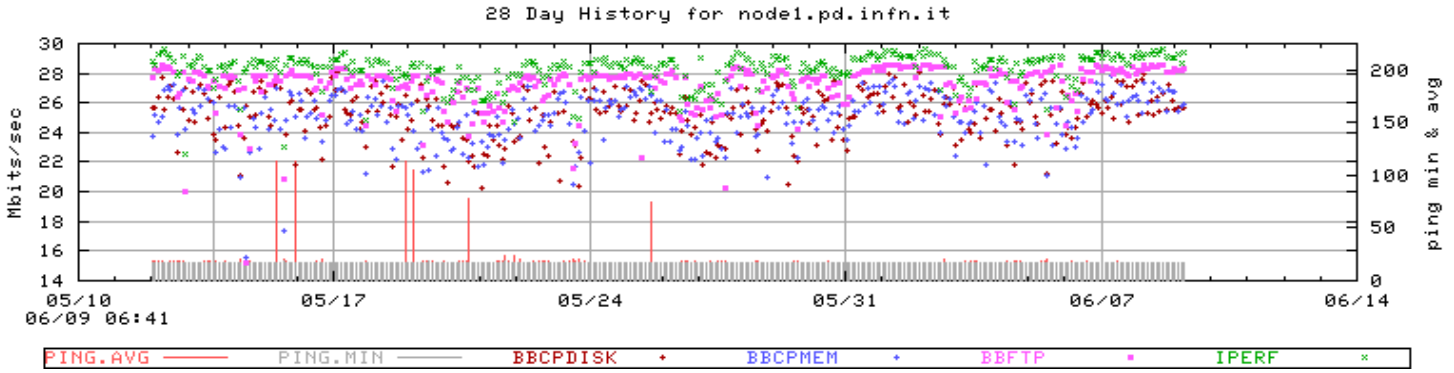
PPST: funziona?

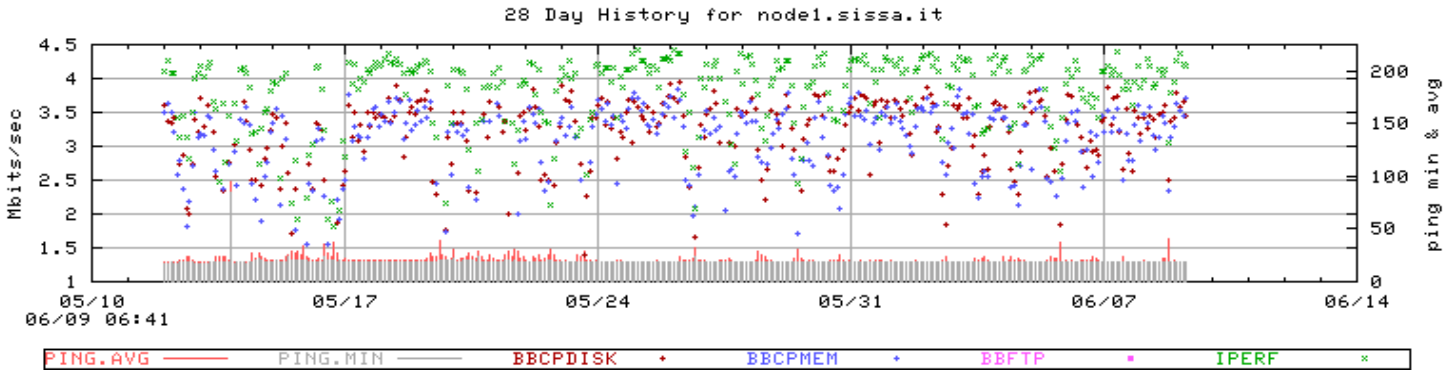
- Cosi' cosi': stima con una certa precisione ed affidabilita' la capacita' massima di un path, ma sbaglia discretamente la stima della banda residua (pur rivelandone variazioni):
    - Treni di misura corti => basso impatto ma bassa statistica, analisi piuttosto difficile (distribuzioni multimodali)
    - Treni di misura lunghi => stima piu' precisa ma maggiore impatto.
- La tecnica pero' e' interessante, e per fortuna ci sono molti tool che funzionano meglio di PPST (poco piu' di un esperimento), sia pure con qualche caveat (ABwE: 40kbits per < 1sec => risultati comparabili con iperf in piu' dell'80% dei casi). Conviene quindi utilizzarli congiuntamente a tool come iperf o similari.

IEPM-BW

- Topologia a stella: un *monitoring node* per ogni rete da misurare; il centro stella controlla tutte le macchine remote (ssh + perl script);
- Parametri misurati: *banda (misura diretta con vari tool e stima), correlazioni, RTT, packet loss*
- Impatto sulla rete: configurabile
- Nodi: PC standard, possibilmente dedicati

IEPM-BW data


28 Day History for node1.pd.infn.it
Mbits/sec
ping min & avg
30
28
26
24
22
20
18
16
14
200
150
100
50
0
05/10
06/09 06:41
05/17
05/24
05/31
06/07
06/14
PING.AVG
PING.MIN
BBCPDISK
BBCPMEM
BBFTP
IPERF


28 Day History for node1.sissa.it
Mbits/sec
ping min & avg
4.5
4
3.5
3
2.5
2
1.5
1
200
150
100
50
0
05/10
06/09 06:41
05/17
05/24
05/31
06/07
06/14
PING.AVG
PING.MIN
BBCPDISK
BBCPMEM
BBFTP
IPERF

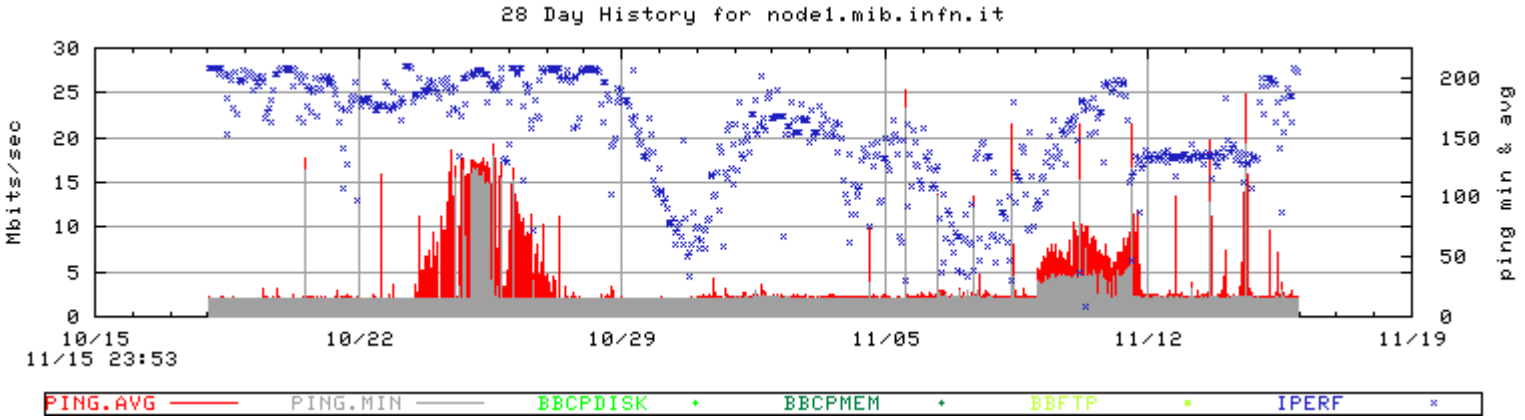
IPM: un embrione di monitoring network per l'INFN (dal 2002)

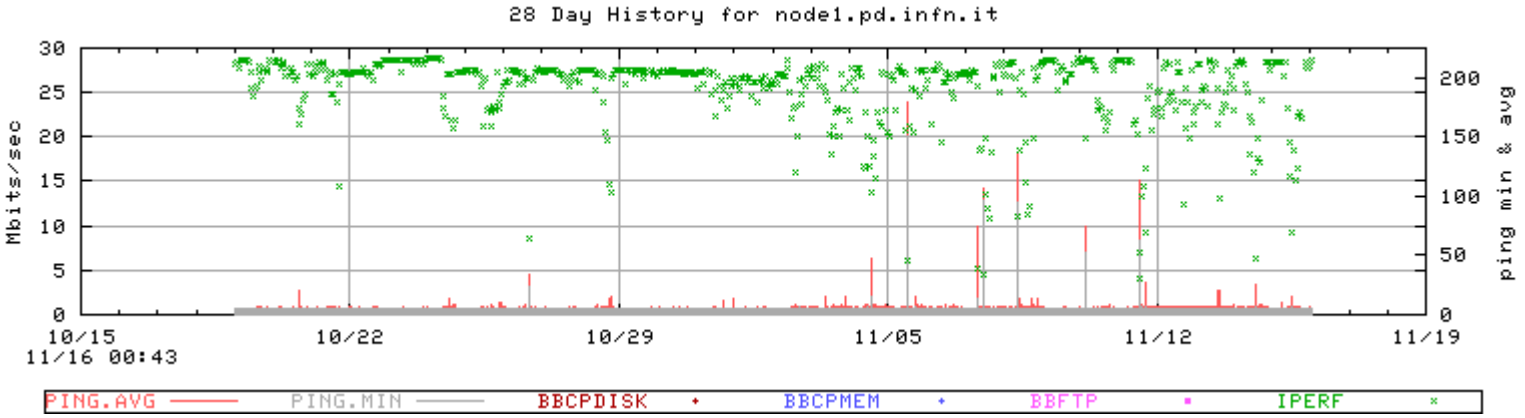
- Struttura e tool mutuati da IEPM-BW, ma su H/W dedicato a basso costo (~ pc industriali):
  - IPMbox piccola: VIA C3 @ 533MHz, 64MB, 3 x 10/100 Mbps, RS232, 1 PCI, mini IDE, SO (RH 7.2) su compact flash da 256MB
  - IPMbox grande: PIII @ 1.2GHz, 128MB, 5 x 10/100, RS232, 2 PCI, mini IDE, 1U, SO (RH 7.2) su compact flash da 256MB
- 1MHz CPU cycle ~ 1Mbits/s (Cottrell et al): potenza sicuramente adeguata per link sino a 100 Mbps (oltre: large pizza box e flussi paralleli)

IPM: nodi e configurazione

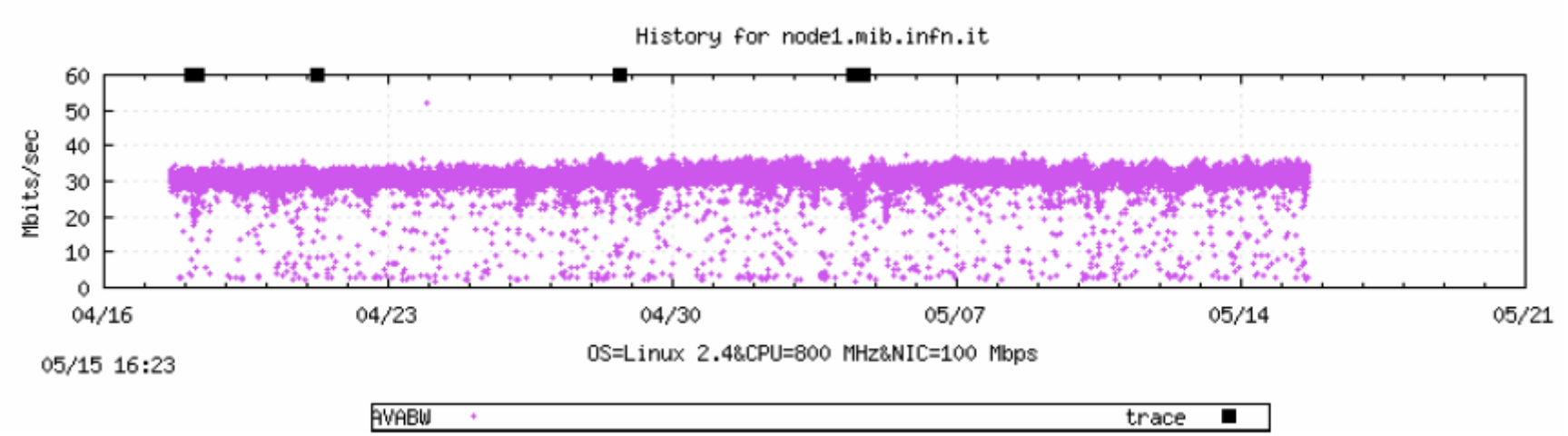
- Nodi centrali: MiB (Bicocca), CERN, Bo
- Nodi periferici: SISSA, Mib, PD, CERN, INAF CT, Uni. Roma III, BO, TS, UD, CILEA, PV
- PD e BO sono nodi bersaglio di misure da SLAC, BNL, FNAL; MiB e TS fanno parte di PingER e RIPE-TTM
- Metriche: banda (iperf, bbcp-disk, -mem, -ftp, patchirp, pathload, AbWE, ...) con cadenza oraria (iperf: 1-20 flussi, 10 sec; possibile scendere a 2-4 senza compromettere l'affidabilita' della stima), latenza, packet loss (ping). Correlazioni tra misure di banda.

IPM data (1)

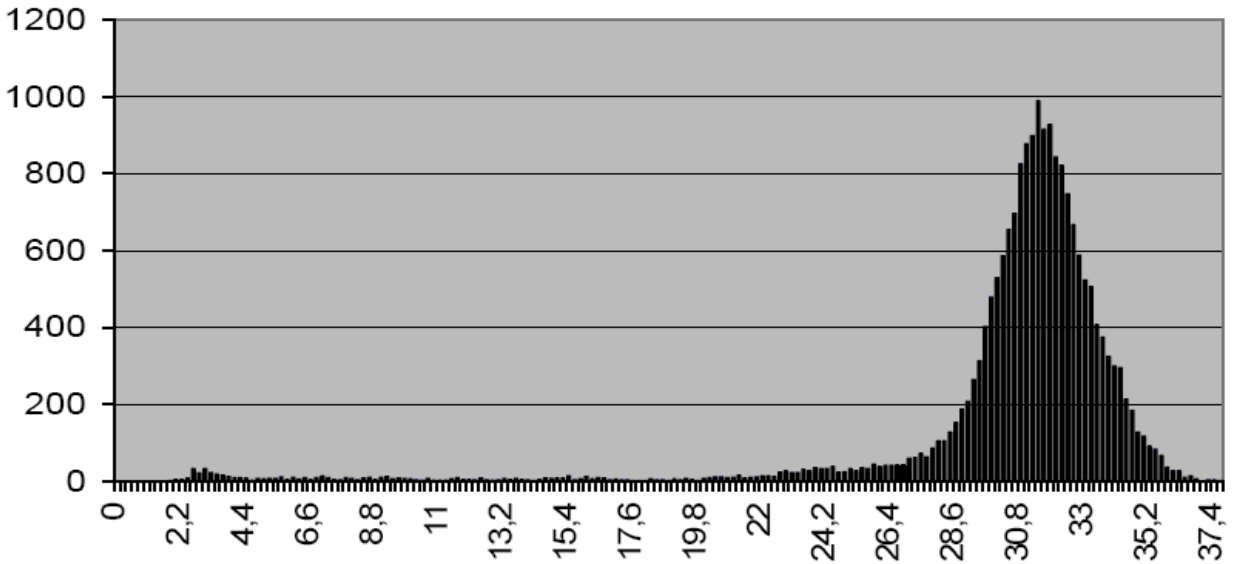

28 Day History for node1.mib.infn.it
Mbits/sec
ping min & avg
30
25
20
15
10
5
0
200
150
100
50
0
10/15
11/15 23:53
10/22
10/29
11/05
11/12
11/19
PING.AVG
PING.MIN
BBCPDISK
BBCPMEM
BBFTP
IPERF


28 Day History for node1.pd.infn.it
Mbits/sec
ping min & avg
30
25
20
15
10
5
0
200
150
100
50
0
10/15
11/16 00:43
10/22
10/29
11/05
11/12
11/19
PING.AVG
PING.MIN
BBCPDISK
BBCPMEM
BBFTP
IPERF

IPM data (2)



Banda disponibile BNL --> ipm02



IPM: ha un futuro ?

- Nato come esperimento di Gr. V INFN (L.Carbone, F.Coccetti, P.Dini, R.Percacci, A.Vespignani), infrastruttura realizzata nel periodo 2003-2004 (inizialmente anche per scopi 'culturali')
- Ha raggiunto il suo scopo? Si – si puo' realizzare e mantenere in operazione una struttura distribuita di misura senza interferire sul funzionamento normale della rete (senza 'disturbare' il management locale) ottenendo comunque molti informazioni sul suo funzionamento
- Scarso successo: poca sensibilita', diffidenza, ...? (un ringraziamento – dovuto - a tutti coloro i quali hanno deciso di ospitare le sonde)
- Possibile evoluzione (in vista dei Tier2?): estensione a piu' siti, potenziamento box, ridefinizione set di misure (es.: iperf -> AbwE, maggior frequenza, ...), creazione di un vero 'Performance DB', ...

Tirando le somme: misurare le rete serve?

- Permette di capire quali siano le reali prestazioni della rete; aiuta a stabilire come raggiungerle.
- Permette di controllare il funzionamento della rete, aiuta ad individuare i problemi ed a risolverli (allarmi, ...)
- Permette di prevedere le prestazioni delle applicazioni di rete, e di progettarle in modo che riescano ad adattarsi automaticamente alle condizioni della rete (server selection, windows, streams, ...)
- Permette di stabilire quali correlazioni esistano tra le varie quantita' misurate ed in che relazione siano queste con le prestazioni percepite (o sperimentate) della rete.
- Le strutture di misura distribuite sono uno strumento a disposizione di tutti per lo sviluppo, la sperimentazione, la validazione di applicazioni di rete (siano esse tool di misura o di produzione).

Sappiamo cosa misurare?

- Non del tutto. La banda residua di un link saturato da un singolo flusso TCP e' nulla. Ma:
  - Se inietto un secondo flusso TCP e' ragionevole misurare un throughput pari a meta' della capacita' del link;
  - Se inietto due flussi TCP otterro' ragionevolmente un throughput pari ad un terzo della capacita' del link;
  - ...
- La banda residua tende quindi realmente a zero solo all'aumentare del numero di flussi.