NetPerF Consulting
more control for better performances !

Come misurare l'end user experience e quali sono le metriche che contano ?

Indice

1. Introduzione

Quante volte c'è capitato di sperimentare noi stessi, direttamente o indirettamente gli effetti di una applicazione lenta o di un servizio inefficiente, utilizzando delle applicazioni via internet, stando di fronte ad uno sportello o concludendo una transazione elettronica ?

Personalmente più di una volta mi è successo di dover rinunciare a concludere l'acquisto di un bene/servizio, vista la lentezza esasperante della interfaccia/applicazione con la quale dovevo interagire, posticipandolo ed in alcuni casi rinunciandovi definitivamente, semplicemente perché è svanito l'attimo in cui eravamo ben predisposti all'acquisto di un nice to have, ma in fondo non così indispensabile.

Ebbene, è estremamente facile pensare subito all'e-commerce, alle transazioni elettroniche, ma in realtà le situazioni in cui l'esigenza di disporre di applicazioni performanti e servizi efficienti è estremamente vasta ed interessa ogni singola azienda e organizzazione che voglia ridurre i propri costi e sprechi, massimizzando gli investimenti nelle aree di effettivo bisogno, aumentando la soddisfazione dei propri utenti finali incrementando la produttività e migliorando la qualità e l'efficienza dei propri servizi.

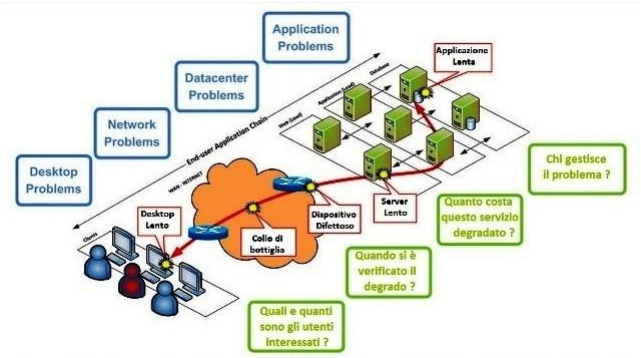


Figura 1

Tutto ciò è possibile solo se il focus viene posto sull'utente, colui il quale utilizza la nostra applicazione o il nostro servizio e non solo su chi tale applicazione o servizio eroga.

E' dunque necessario iniziare a misurare l'end user experience, ovvero la percezione che hanno gli utilizzatori, ovunque essi siano, di una applicazione/servizio, utilizzando metriche oggettive che ci permettano di comprendere appieno il fenomeno che vogliamo analizzare.

2. Come misurare l'End User Response Time ?

Monitor attivo o passivo ? Quale soluzione scegliere ?

Soluzioni attive o di active test, sono da utilizzarsi per controllare la disponibilità di una applicazione o di un servizio, ma certamente non sono idonee a dare una rappresentazione veritiera di quanto viene percepito dagli utilizzatori, che accedono alla applicazione/servizio con tecnologie diverse e dai posti più disparati.

L'obbiettivo non deve essere quello di misurare le prestazioni di una applicazione o di un servizio accedendovi ripetutamente, eseguendo le medesime operazioni dall'interno della rete stessa od in

prossimità della stessa, quanto piuttosto misurare ciò che percepiscono gli utenti che per accedervi attraversano innumerevoli dispositivi e spesso utilizzano molteplici tecnologie e modalità di accesso la cui banda passante è decisamente limitata.

Quindi, se siamo interessati a conoscere la reale end user experience, il monitor deve essere di tipo passivo, ovvero dovremo misurare le performance percepite dagli utenti analizzando il traffico dati generato tra gli utenti medesimi e le applicazioni erogate dal data center.

3. Come intercettare il traffico da analizzare ?

Affinché si possa analizzare il traffico generato tra gli utenti e le applicazioni, è necessario prima di tutto intercettarlo nel modo più opportuno, ovvero utilizzando la tecnica migliore che eviti la perdita di pacchetti con la conseguenza di perdere informazioni utili per una misura corretta scevra da errori.

La possibilità di perdere pacchetti, dipende da molteplici fattori, quali : velocità del media utilizzato 1Gigabit/10Gigabit, rate del traffico, capacità di analisi del sistema di misura e tecnica utilizzata.

Il sistema preposto all'analisi dei dati è solitamente una appliance (un sistema custom particolarmente efficiente e performante) o un server (linux o windows) corredato di apposito software, su cui deve essere convogliato il traffico da analizzare, pertanto la quantità di traffico (rate) non potrà essere superiore alla capacità reale di analisi del sistema stesso (esempio – 1Gbps).

Qualora il rate del traffico da analizzare fosse superiore alla capacità di analisi del sistema di misura si dovranno utilizzare tecniche atte a ridurre la quantità di traffico stessa o limitare l'inoltro di tale traffico, verso il sistema di misura, al solo traffico relativo alle applicazioni che si vogliono analizzare.

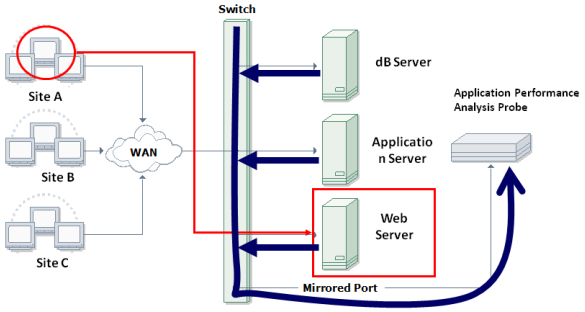


Figura 2

Il modo più semplice ed economico non è detto che sia sempre il migliore, dipende dalla quantità di traffico (rate) in gioco; ovvero le tecniche di SPAN/Mirror che copiano traffico da una o più porte o da una o più VLAN di uno switch su di una o più porte di output su cui è attestato il sistema di misura, sono generalmente delle valide soluzioni per le analisi di front-end ma spesso non sono adeguate per le analisi di back-end dove i rate in gioco sono decisamente superiori.

In quest'ultimo caso, si dovrà ricorrere all'utilizzo di TAP e Media Filtering Aggregator, dispositivi in grado di spillare il traffico da un link in rame o fibra, di riceverlo da una SPAN Port e di replicarlo in output su una o più porte aggregandolo e filtrandolo a seconda delle necessità.

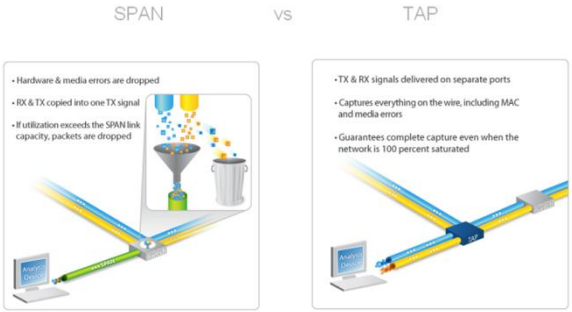


Figura 3

Si tratta di dispositivi particolarmente intelligenti e molto versatili, alcuni dei quali totalmente passivi o attivi (dal punto di vista elettrico); il mercato attuale offre innumerevoli soluzioni.

4. Dove intercettare il traffico da analizzare ?

Dal punto di vista delle soluzioni preposte alla misura della end user experience, esistono due possibili approcci : soluzioni distribuite e soluzioni centralizzate.

Le soluzioni distribuite prevedono il deployment di probe/appliance in ogni sito dove esistono degli utilizzatori delle applicazioni che si vogliono monitorare, ciò permette di poter essere molto vicino ai client che generano il traffico, ma comporta costi di acquisizione e gestione particolarmente elevati, ad esempio nel caso in cui semplicemente avessimo 500 utenti distribuiti su 10 siti, ebbene si dovrebbero installare 10 probe/appliance, uno per ogni sito.

L'esigenza di doversi dotare di una soluzione distribuita è reale solo nel caso in cui l'applicazione che si vuole monitorare è una applicazione VoIP, nel caso cioè in cui si voglia valutare la end-to-end VoIP Call Qulity, ma nel caso delle applicazioni transazionali basate su TCP, non vi è alcuna esigenza di dover distribuire probe/appliance nei siti remoti, basta utilizzare una soluzione centralizzata.

Il mercato ha premiato i vendor che negli ultimi anni hanno proposto soluzioni centralizzate.

Le soluzioni centralizzate consistono nel deployment di una o più appliance nel data center, a seconda del livello di analisi richiesto, ovvero se ci si deve limitare ad analizzare solo il front-end oppure è anche richiesta l'analisi a livello di back-end, come mostrato in figura 4.

Figura 4

Questa figura mostra una classica applicazione n-tier, cioè a più livelli, dove il livello di front-end è rappresentato dalla applicazione web.

Esistono per la verità, soluzioni che anziché adottare appliance da posizionare nel data center, prevedono il deployment di agent da installarsi direttamente sui server.

Questo tipo di soluzione, non è in grado di evidenziare la end user experience, ma semplicemente consente una analisi di maggiore dettaglio di quanto accade al server su cui viene installato.

Seppur interessante, per le capacità di analisi di dettaglio offerte, risulta del tutto evidente l'invasività della stessa ed il rischio, non trascurabile, di possibili conflitti quanto già installato nonché l'esigenza di doverne costantemente allineare le versioni alle diverse versioni di OS e moduli applicativi.

Quindi in conclusione, una soluzione centralizzata basata su una o più appliance installate nel data center, vicino ai server, è senza alcun dubbio la scelta migliore, visto oltretutto che in questo caso è in grado di analizzare in modo passivo, non invasiva, il traffico delle applicazioni proveniente da tutte le nostre sedi periferiche, anche per connessioni che avvengono via internet o mobile, con enormi benefici in termini di costi e gestione, senza peraltro precluderci alcunché o limitarci nella analisi delle end user experience.

5. Cos'è il tempo di risposta ?

Si consideri un utente (un PC client) al lavoro su un'applicazione che comunica con un server attraverso una rete. Il client sottopone una richiesta al server, e quest'ultimo risponde con uno o più pacchetti. In generale, una transazione (per esempio l'invio di un ordine o l'effettuazione di una query) può consistere in una serie di richieste da parte del client e delle corrispondenti risposte del server. Il tempo trascorso da quando il client invia la sua richiesta sino a quando riceve l'ultimo pacchetto della relativa risposta è ciò che viene definito come Tempo Totale di Risposta. A formare questo Tempo Totale di Risposta contribuiscono i comportamenti di reti, server e applicazioni.

E' necessario pertanto che la soluzione di Application Performance Analysis adottata sia in grado di analizzare ogni singola transazione nel dettaglio al fine di identificare le singole metriche che contribuisco a formare il tempo complessivo percepito dall'utente, ovvero l'EURT (End User Response Time).

6. Cos'è l'EURT (End User Response Time)

L'appliance posizionata nel data center intercetta ed analizza tutte le richieste provenienti dai clients e le risposte provenienti dai server, iniziando a tracciarle sin dal primo pacchetto contenente il TCP Syn, utilizzato come una sorta di trigger per iniziare a collezionare tutte le statistiche associate alle metriche che l'appliance è in grado di monitorare. La figura 5 esemplifica il flusso di una applicazione web.

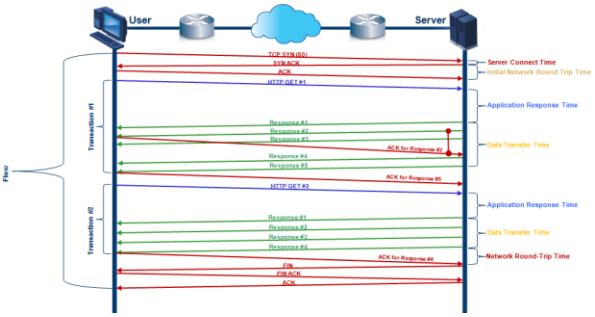


Figura 5

Evidenziando i singoli macro componenti, che costituiscono l'EURT, ovvero ART (Application Response Time), DTT (Data Transfer Time) e NRTT (Network Round Trip Time) a connessione TCP instaurata, dopo la fase di hand-shacking (TCP-SYN, TCP-SYNACK, TCP-ACK).

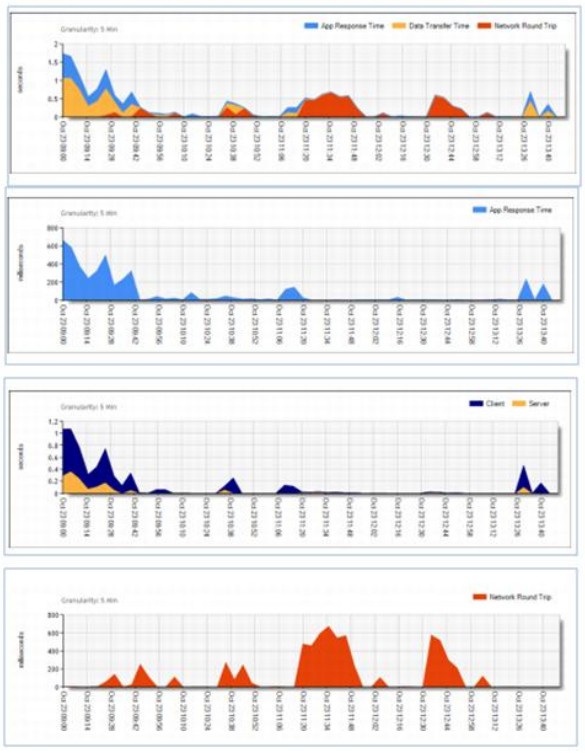
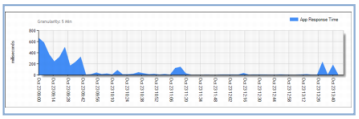


Figura 6

7. Le principali metriche in gioco

ART 

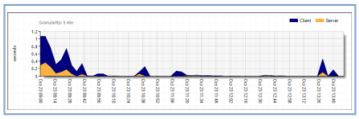
Rappresenta il tempo che l'application server impiega per rispondere ad una richiesta di un client, con il primo pacchetto contenente dati

A cosa serve ?

ART può essere utilizzato per monitorare le performance di una applicazione, identificare dei server sovraccaricati ed indica la client experience quando viene utilizzato come parte della metrica EURT (End-User Response Time)

Da cosa è impattato ?

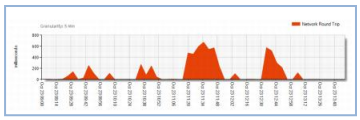
- dai processi di back-end a valle, dal next tier a valle
- dalle prestazioni del Server e dal suo carico
- da uno specifico utente o transazione

DTT 

Rappresenta il tempo complessivo impiegato per trasmettere completamente una risposta ad una richiesta, misurato partendo dal primo all'ultimo pacchetto di una risposta. Quando utilizzato come parte della metrica EURT, il DTT è misurato come il valore medio combinato del tempo di trasferimento dei dati della richiesta del client e dei dati della risposta del server.

Da cosa è impattato ?

- dalle prestazioni di : Client, applicazione, network, o server
- da condizioni normali, causato da una inusuale transazione molto grande
- da una applicazione inefficiente

NRT 

Rappresenta il tempo richiesto ad un pacchetto dati della applicazione, per attraversare la rete da una specifica sorgente verso una specifica destinazione e consentire alla risposta di ritornare .

Questo comprende il tempo di transito dall'elemento di processo di rete alla rete , come pure il TCP delay.

Da cosa è impattato ?

- da un sovraccarico o congestione sulla network in upstream
- da un sovraccarico del upstream client

Alcune metriche sono estremamente esemplificative nel localizzare il problema, è il caso ad esempio dell'ART (Application Response Time), altre come il DTT (Data Transfer Time) o il NRTT (Network Round Trip Time) necessitano di essere correlate con altre metriche, inerenti il traffico in rete, il TCP layer di trasporto, le performance dei router, dei server ed altro.

La condizione necessaria per aver successo è quella di disporre di una soluzione integrata che sia in grado di correlare metriche acquisite con modalità diverse, eventualmente anche in punti diversi, ma sincronizzate utilizzando lo stesso sample time.

8. Correlazioni e statistiche

Le metriche : EURT,ART,DTT e NRTT sono senza alcun dubbio le principali ma devono essere correlate da altre affinché si possa ottenere una visone corretta di quanto si sta osservando; è questo il caso del DTT e del NRTT.

Questo significa inevitabilmente che la misura della End User experience necessita della integrazione della network performance analysis, in particolare occorre una analisi complessiva del traffico in rete; pertanto la soluzione di traffic analysis deve essere integrata con la soluzione di application performance analysis, devono cioè avere la medesima granularità affinché le informazioni siano realmente fruibili ed abbiano un significato, come evidenziato dalla Figura 7.

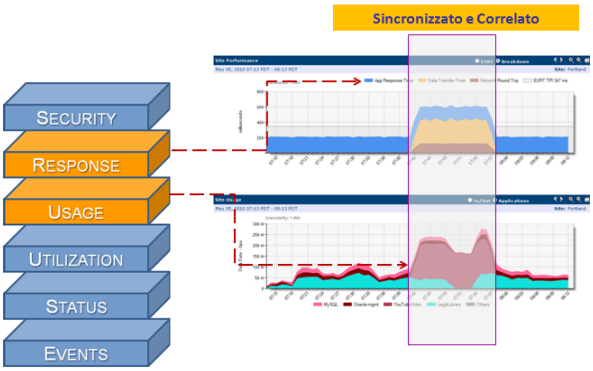


Figura 7

Perché tutto questo ?

I fattori che spesso limitano le performance di una applicazione, sono : bandwidth, latency, jitter e packet loss, in particolare nel caso di applicazioni transazionali, bandwidth e latency sono quelli che giocano un ruolo determinante, che trovano una diretta correlazione con le metriche : DTT e NRTT.

La latenza (latency) è il tempo complessivo necessario per spostare un bit da una location ad un'altra.

La latenza è impattata da diversi fattori, tra i quali :

- La distanza tra le locations
- Il percorso di Routing
- Il tempo di transito Router/switch
- La congestione e gestione delle code di Router/switch
- Un utilizzo eccessivo del link di connessione con il Carrier

Pertanto, applicazioni le cui performance sono limitate dalla latenza, hanno queste caratteristiche :

- **Molti "turns" per transazione**. Una applicazione che usa un eccessivo numero di "turns" avrà gli effetti della latenza moltiplicata per il numero di "turn" richiesti

- **Small packet sizes**. Pacchetti di dimensione piccola, come risultato della frammentazione dei dati della applicazione, significa un gran numero di frame. L'effetto della Latenza è moltiplicato dal numero di frames inviate

- **Il limite imposto dalle capacità della Window Size** del protocollo impone il rapporto dei "turn", come nel caso dei pacchetti piccoli, il risultato è quello di avere un gran numero di frame moltiplicando l'effetto della Latenza.

- **Un elevato overhad del protocollo** per trasportare i dati della applicazione riduce lo spazio disponibile per il payload della applicazione, per cui il risultato è che servono più frames per trasportare tutti i dati della applicazione, moltiplicando l'impatto della Latenza.

- **Il rate, inferiore alla banda passante disponibile**. Anche se il bandwidth (larghezza di banda) è insufficiente, aumentarla non significa necessariamente migliorare le performance , se il flusso è influenzato da altri vincoli di latenza.

Le applicazioni che sono invece limitate dalla banda passante disponibile, hanno le seguenti caratteristiche :

- **Pochi "turns" per transazione**

- **Windows size di grandi dimensioni**

- **Poco sensibili al tempo**

Pertanto le principali metriche devono essere integrate da informazioni che siano in grado di descrivere assai bene lo stato di salute del TCP layer : TCP retransmission, TCP zero windows, TCP resets, TCP out of orders evidenziando naturalmente la direzione degli stessi, ovvero se tra client e server o viceversa.

In modo del tutto analogo è necessario che la soluzione adottata sia anche in grado di evidenziare : SCT (Server Connection Time), INRTT (Initial Network Round Trip Time) e CST (Connection Setup Time), come evidenziato dalla figura 5.

9. La granularità necessaria

Il mercato offre soluzioni in grado di evidenziare metriche e statistiche con granularità : 1min., 2min., 5min, 15min., …

Certamente poter disporre di una soluzione che sia non solo in grado di monitorare tutte le metriche che servono davvero, che le correli tra loro, che le correli con i dati di network performance analysis, ma lo faccia anche con la massima granularità possibile, significa poter disporre di un sistema estremamente efficace nell'individuare ed isolare il reale problema, separandolo tra client, rete, server ed applicazione.

Le statistiche sono gioco forza aggregate e mediate rispetto al sample time scelto, il cui valore minimo come detto può essere al massimo di un minuto, per questo motivo le statistiche di monitor devono poi essere affiancate dalle capacità di analisi intrinseche della soluzione adottata, necessarie per poter fare uno zoom all'interno delle transazioni di un singolo sample time, ad esempio di 1 minuto.

10. Analisi Layer 4 o superiore ?

Al fine della corretta individuazione di una tendenza, il limitarsi al TCP Layer 4 è sufficiente, certamente non lo è più nel momento stesso in cui si voglia individuare esattamente gli eventuali application statement responsabili dell'incremento di una metrica come l'ART.

In altre parole, il fatto di aver individuato 'ART come responsabile del degrado dell'EURT percepito da un utente, è in alcuni casi non sufficiente, in quanto spesso si richiede anche di conoscere esattamente il motivo, la causa reale di detto degrado da parte dell'ART.

Ciò richiede soluzioni in grado di analizzare ogni singola transazione di ogni singolo client ad un livello applicativo, superiore al TCP Layer 4, con capacità di application analysis più in profondità al fine di interpretare lo statement associato alla request/response, senza dover necessariamente ricorrere alla analisi di una packet trace.

La packet trace è certamente ben accetta, ma deve essere invocata solo come ultimo strumento in quanto tipicamente complessa da interpretare e quindi time consuming, molto meglio affidarsi a soluzioni di application statement recognition, peraltro in grado di evidenziare istantaneamente variazioni rispetto ad una baseline acquisita nel tempo.

Tutto ciò si traduce in facilità di utilizzo ed immediatezza nel trovare le risposte che si cercano giornalmente.

11. n-Tier Analysis

Le applicazioni sono oggi sempre più complesse e stiamo assistendo a numerosi cambiamenti, in particolare :

a. ad un massivo utilizzo del web, quella che in gergo si chiama webificazione delle applicazioni
b. Server consolidation
c. Data Center consolidation e single hosting
d. Virtualizzazione
e. Cloud computing

Ma ciò che va sottolineato, è che sempre più spesso, quando le persone accedono ad una applicazione, lo fanno attraverso una WAN e non in LAN, quindi si è sempre più distanti dalla applicazione !

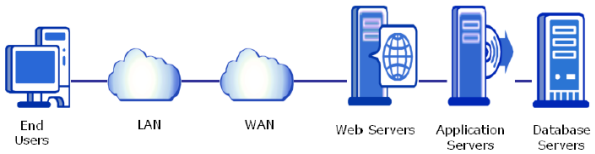


Figura 8

Una soluzione di application performance analysis deve essere pertanto in grado non solo di analizzare l'End User Response Time , a livello di front end, ma deve all'occorrenza essere capace di correlare quanto percepito dagli utenti, con quanto accade a livello di back-end onde ottenere le informazioni necessarie per isolare ed identificare la reale causa di un problema.

Siccome stiamo parlando di soluzioni di monitoring passivo, non invasive, che non utilizzano agents, ne sui client ne tantomeno sui server, queste soluzioni devono essere caratterizzate da funzionalità, granularità, timing clock particolarmente spinte, affinché analizzando tutto il traffico tra client e front-end server, front-end server e application server e tra application server e db server, siano in grado di associare e correlare gli eventi, semplicemente utilizzando la medesima base temporale.

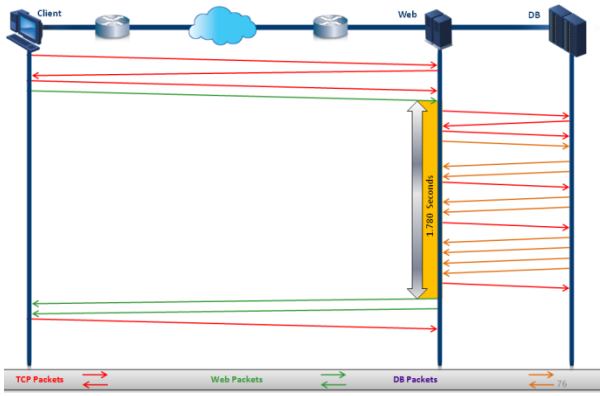


Figura 9

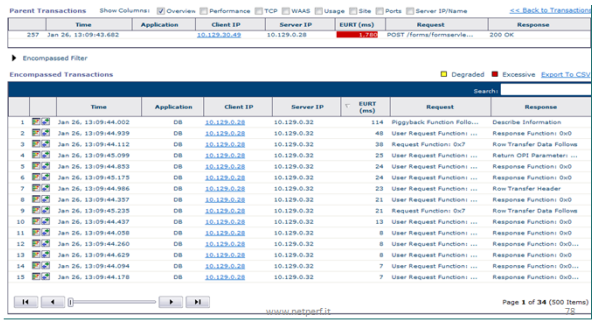


Figura 10

12. Summary

Da quanto sin qui evidenziato, possiamo elencare le caratteristiche che dovrebbe avere una soluzione che sia in grado di misurare la End User experience :

a. Monitoring passivo
b. Architettura server-side (Datacenter)
c. No client side agents
d. No server side agents
e. Facile da implementare e manutenere
f. Misura tutto il traffico reale degli utenti, proveniente da tutte le ubicazioni

g. E' in grado di scomporre istantaneamente l'EURT (End User Response Time) nelle sue componenti principali : ART,DTT,NRTT
h. Fornisce l'analisi del traffico server–to-server , n-tier analysis
i. Correla le diverse metriche, anche di network performance analysis, utilizzando la massima granularità possibile, 1 minuto
j. Analizza in profondità le applicazioni, non solo a livello L4
k. Analizza ogni singola transazione, con dati di dettaglio e non solo summary
l. Non richiede di dover analizzare necessariamente una packet trace, ma deve essere in grado di fornire dei dati di sintesi degli statements associati alla transazione

13. Domande da porsi, prima di scegliere una soluzione di APM

a. E' in grado di monitorare tutte le applicazioni (TCP-UDP)?
b. E' in grado di scomporre le prestazioni di applicazioni n-Tier ?
c. E' in grado di scomporre un problema di prestazioni EURT, nelle sue componenti, ovvero : ART,DTT,NRTT ?
d. E' in grado di analizzare ogni singola transazione di ogni singolo client, o fornisce solo dati aggregati ?
e. Può analizzare il tempo di risposta di applicazioni su un periodo di tempo esteso ?
f. Misura il tempo di risposta reale o qualche sorta di simulazione sintetica ?
g. Cosa deve essere implementato, dove e come verrà gestito ?
h. Quanto è difficile ottenere informazioni utili ?
i. E' in grado di investigare problemi in modo autonomo, documentando tutti i processi ?
j. Qual'è il modello di licensing ?
k. Qual'è il *vero costo totale* della soluzione completa ?

Autore : Maurizio Malinconi

NetPerF Consulting
more control for better performances !

Via A. Diaz, 30
20035 Lissone (MB)
Tel. +39 02 40047334
Fax +39 039 2781283
Email info@netperf.it
Web www.netperf.it