Analisi e sviluppo di nuove tecniche per l'estrazione di informazioni da grandi moli di dati provenienti dal web

Giuseppe SANTOMAURO

Tutor: Ing. **Giovanni Ponti**
(DTE-ICT-HPC, ENEA C.R. Portici)

Agenzia nazionale per le nuove tecnologie,
l'energia e lo sviluppo economico sostenibile

7° Borsisti Day

20/01/2016

Roma – Consortium GARR

Borse di Studio
Orio Carlini
Consortium GARR

Borse di Studio Orio Carlini Consortium GARR

Attività svolte

Predisposizione di un ambiente per il web crawling

Web Data Retrieving: **esplorazione dei contenuti in una rete in maniera sistematica e automatizzata al fine creare archivi di dati web.**

Operazione eseguite:

- Installazione di strumenti di web crawling sul cluster HPC CRESCO presente nel Centro Ricerche di Portici;
- Integrazione in ENEA-GRID;
- Definizione di un'infrastruttura hardware ad-hoc per il crawling e confinamento delle risorse;
- Tuning dei parametri del crawler;
- Pianificazione ed esecuzione di sessioni di web crawling in diverse condizioni e con di diversa durata.

Borse di Studio Orio Carlini Consortium GARR

Tempistiche

1) Prima fase [~2 mesi]:
- Studio e individuazione delle metodologie per il web crawling;
- Analisi e individuazione dei prodotti software.

2) Seconda fase [~4 mesi]:
- Studio dell'infrastruttura ENEA-GRID/CRESCO;
- Individuazione del tipo e della quantità delle risorse fisiche da impiegare nell'attività di crawling.

3) Terza fase [~4 mesi]:
- Installazione, configurazione e prime esecuzioni di test dei prodotti software;
- Tuning dei parametri e individuazione di configurazioni ottimali.

4) Quarta fase [~2 mesi]:
- Esecuzione di crawling di grandi dimensioni;
- Analisi prestazionale dei risultati e delle perfomance.

Prima Fase

Strumenti e metodologie per il web crawling

Problematiche e Normative

- Ricerca delle best practices al fine di evitare eccessivi sovraccarichi della rete e/o di suoi utilizzi in modo improprio;
- Individuazione delle leggi che regolano il processo di crawling al fine di rispettare i diritti di privacy e/o copyright.

Prodotti

- Utilizzo di soluzioni open source;
- Crawling standard (download pagine web);
- Crawling avanzato (download + parsing + strutturazione + preanalisi)

Problematiche e Normative

Denial of Service:

- Rallentamento dell'attività di un web server causata da una ripetuta richiesta di pagine oppure dall'esaurimento delle risorse di banda della rete;
- Può essere di due tipi: accidentale o intenzionale.

Privacy:

- I contenuti sul Web sono di dominio pubblico;
- Informazioni aggregate su larga scala e su molte pagine.

Copyright:

- Molti motori di ricerca emulano l'attività di **Internet Archive**:
  - Rispetto del protocollo *Robots Exclusion Standard*;
  - Richiesta di rimozione dall'archivio.
- In Italia:
  - Legge n. 633 del 22 aprile 1941 (Protezione del diritto d'autore).

Prodotti individuati

Crawling standard

- **Heritrix;**
- **Crawler4j;**
- **Nutch;**
- **BUbiNG.**

Crawling con preanalisi dei dati

- **Scrapy;**
- **OpenWebSpider;**
- **OpenSearchServer.**

Scelta dei prodotti

Criteri di selezione:

- Possibilità di integrazione nell'infrastruttura ENEA-GRID/CRESCO;
- Flessibilità sulla configurazione del prodotto;
- Strumenti e interfacce per il monitoring dell'esecuzione;
- Formato di archiviazione dei dati web;
- Capacità di garantire le migliori performance.

Nutch

BUbiNG

Prodotti installati

Nutch:

- Basato su *Lucene* e *Java;*
- Codificato interamente in Java, ma i dati vengono scritti in formati indipendenti dal linguaggio;
- Architettura altamente modulare, che consente agli sviluppatori di creare plug-in per media-type parsing, data retrieval, querying and clustering;
- Accessibile da terminale.

BUbiNG:

- E' costruito sull'esperienza decennale di **UbiCrawler** (Università Statale di Milano);
- E' scritto in Java;
- Singoli agenti possono eseguire il crawling su diverse migliaia di pagine al secondo rispettando i vincoli di politeness;
- La distribuzione dei job è basata sui moderni protocolli ad alta velocità così da raggiungere un elevato throughput;
- Permette l'esecuzione simultanea di più agenti che possono comunicare tra di loro attraverso la libreria JGROUP;
- Le opzioni di configurazione sono passate su un file e molti di questi parametri possono essere modificati a runtime attraverso l'interfaccia JMX;
- L'output viene salvato in file *.warc*.

Seconda Fase

Infrastruttura hardware

Studio dell'infrastruttura di ENEA-GRID

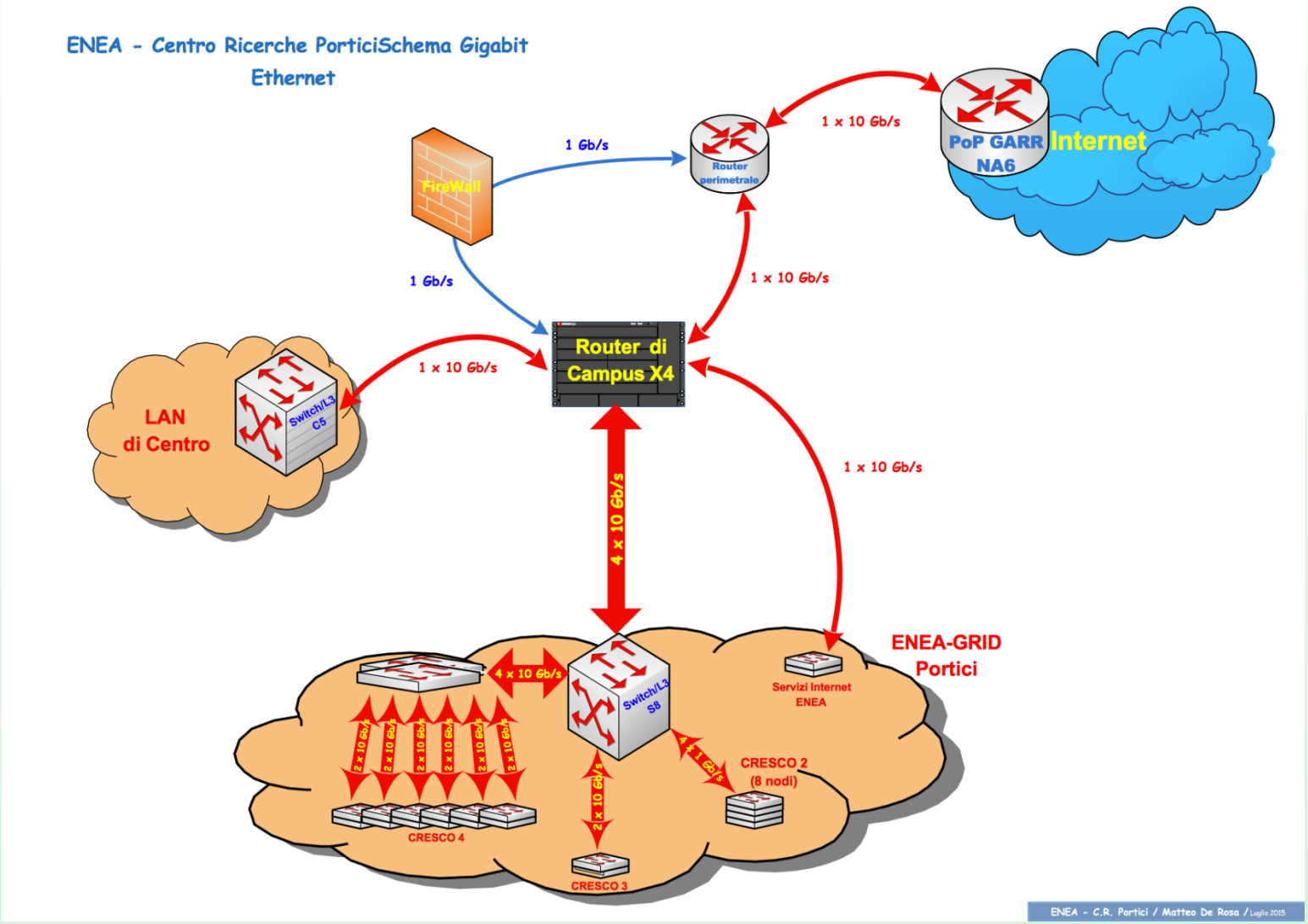
- Condizioni di utilizzo di accesso a CRESCO;
- Sottomissione di job;
- Scheduler delle risorse LSF;
- Configurazione di job paralleli.

Individuazione dell'architettura computazionale e di rete

- Scelta delle Risorse di calcolo;
- Studio dell'infrastruttura di Rete del C.R. Portici e di CRESCO;
- Confinamento delle Risorse.

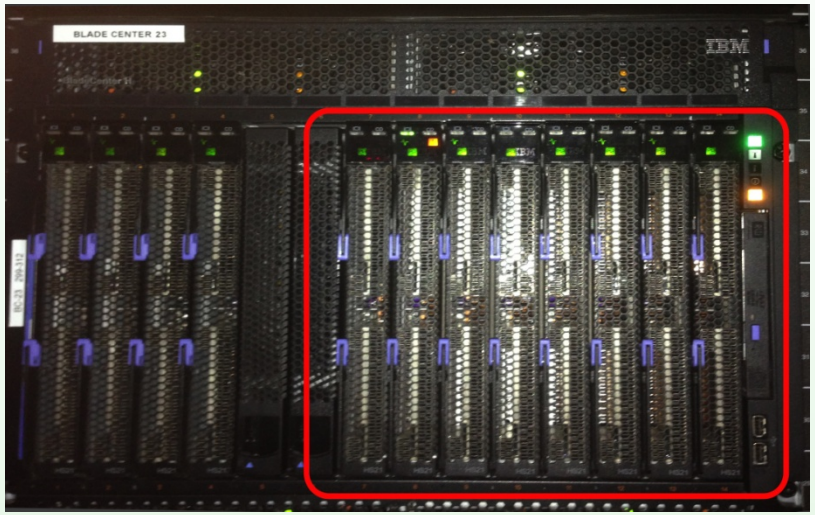
Borse di Studio
Orio Carlini
Consortium GARR

Schema Rete



Borse di Studio Orio Carlini
Consortium GARR

Risorse Hardware



8 NODI della sezione *CRESCO 2*:

- **Processore**: 2 Xeon Quad-Core Clovertown E5345;
- **RAM**: 16 GByte;
- **Clock**: 2.33GHz/1333MHz/8MB L2 state.

Switch *Enterasys S-Series S8*:

- **Chassis Slots**: 8;
- **Ports:** 576 a 1 Gbps o 128 a 10 Gbps
- **System Switching Capacity**:1.28 Tbps;
- **System Switching Throughput**:960 Mpps;

Risorse Hardware

Router di Campus *Enterasys Matrix X4*:

- **Chassis Slots**: 4;
- **Ports:** 768 a 1 Gbps o 48 a 10 Gbps
- **System Switching Capacity**: 640 Gbps;
- **System Switching Throughput**: 119 Mpps.

2 Firewall *Secure Firewall 2150*

- **Packet Filtering Throughput:** 3.1 Gbps;
- **Concurrent Connections:** 1,600,000

Router Perimetrale *Cisco 7606*:

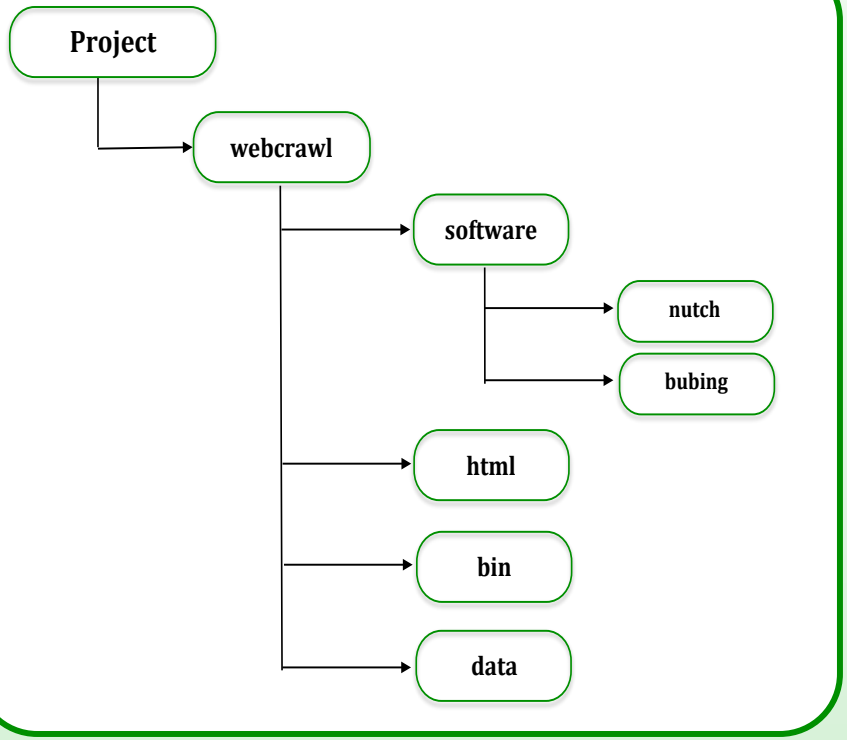
- **Distributed forwarding rate:** up to 240 Mpps;
- **Total throughput:** 480 Gbps.

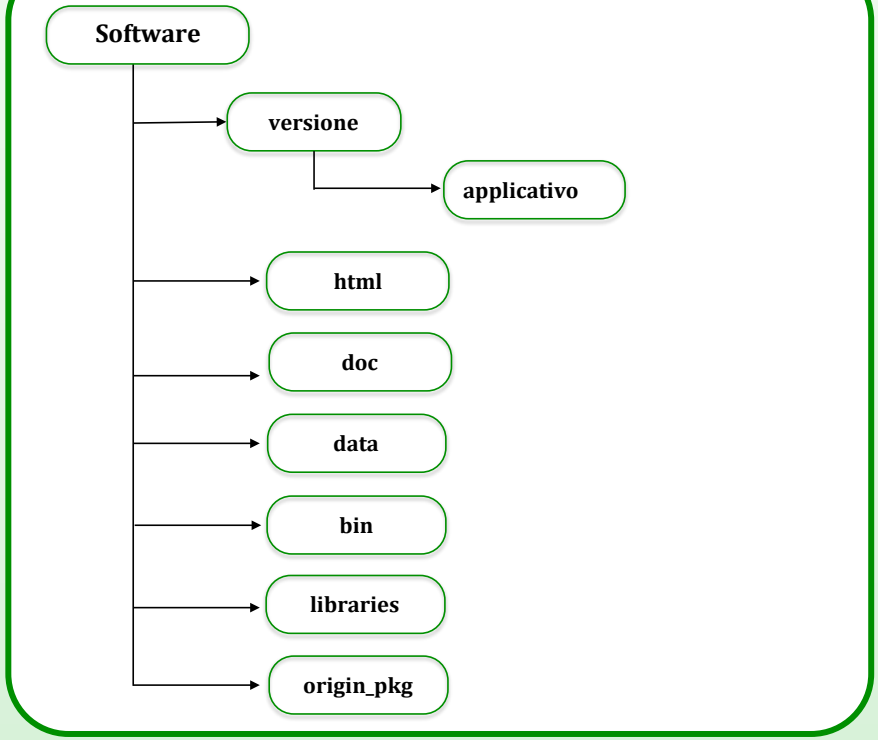
Terza Fase

Installazione software

Creazione sull'area progettuale di ENEA-GRID del volume ***webcrawl.***

Installazione di due crawler: ***Nutch*** e ***BUbiNG.***





BUbiNG

Configurazione e esecuzione software

- **Settaggio del sistema di filtro**: vengono decise, attraverso un insieme di regole logiche, quali pagine web saranno prese in considerazione e quali no;
- **Allocazione della memoria**: vengono settate le quote destinate alle varie memorie che BUbiNG utilizza durante l'esecuzione;
- **Settaggio del numero di threads**: per ogni azione dell'istanza Java attivata vengono stabiliti quanti processi saranno destinati per essa;
- **Scelta del seme**: viene fornita una lista di indirizzi web dalla quale il crawler inizia a scaricare.
- **Scripting:** viene sviluppato un insieme di codici per l'esecuzione di più agenti su più macchine.

Borse di Studio Orio Carlini Consortium GARR

Tuning dei parametri

	8 Agenti	16 Agenti	32 Agenti	64 Agenti
numero di nodi	8	8	8	8
numero di processori per nodo	8	8	8	8
RAM disponibile per nodo in GB	16	16	16	16
numero di agenti per nodo	1	2	4	8
thread stack size	256K	256K	256K	256K
heap memory iniziale in MB per agente	12000	6000	3000	1500
heap memory massima in MB per agente	12000	6000	3000	1500
maxUrlsPerSchemeAuthority per agente	200	200	200	200
parsingThreads per agente	8	4	2	2
dnsThreads per agente	36	36	36	36
fetchingThreads per agente	512	256	64	32
scheduleFilter	.it/	.it/	.it/	.it/
numero di url iniziali per agente	8	4	2	1
schemeAuthorityDelay	10s	10s	10s	10s
ipDelay	2s	2s	2s	2s
maxURLs	500M	250M	64M	32M
bloomFilterPrecision	1,00E-008	2,00E-008	1,00E-008	1,00E-008
socketTimeout	60s	60s	60s	60s
connectionTimeout	60s	60s	60s	60s
fetchDataBufferByteSize	200K	200K	200K	200K
cookiePolicy	compatibility	compatibility	compatibility	compatibility
cookieMaxByteSize	500	500	500	500
robotsExpiration	1h	1h	1h	1h
responseBodyMaxByteSize	2M	2M	2M	2M
digestAlgorithm	MD5	MD5	MD5	MD5
workbenchMaxByteSize per agente	512Mi	512Mi	512Mi	256Mi
urlCacheMaxByteSize per agente	1Gi	1Gi	1Gi	512Mi
sieveSize per agente	128Mi	128Mi	128Mi	64Mi
parserSpec	HTMLParser(MD5)	HTMLParser(MD5)	HTMLParser(MD5)	HTMLParser(MD5)
keepAliveTime	1s	1s	1s	1s
tempo previsto di esecuzione in sec	900	900	900	900
tempo effettivo massimo di esecuzione in sec	1029	1054	1029	1023
totale materiale scaricato di warc in MB	140296,00	152191,00	106973,89	93130,60
velocità media dowload in MB/s	**136,34**	**144,39**	**103,96**	**91,04**
velocità media dowload in MB/s per agente	17,04	9,02	3,25	1,42
velocità media dowload in Mb/s	1090,74	1155,15	831,67	728,29
velocità media dowload in Mb/s per agente	136,34	72,20	25,99	11,38

Quarta Fase

Esecuzione di crawl di grandi dimensioni

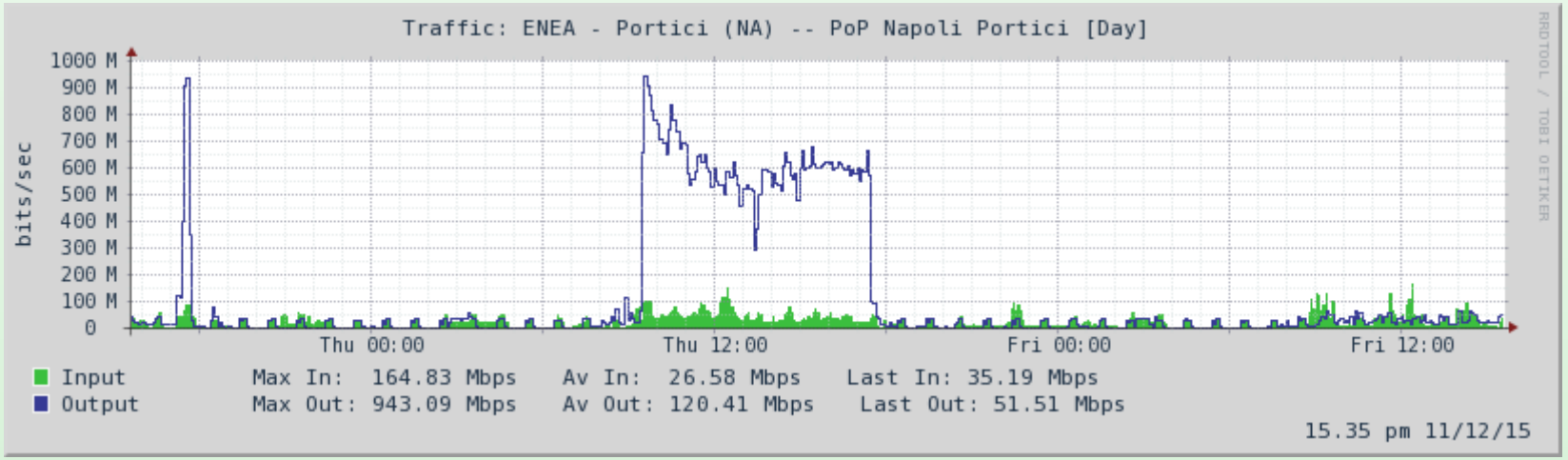
- **Periodo:** sono stati scelti giorni e orari con modesto traffico;
- **Avviso di esecuzione:** segnalazione ai tecnici della rete di attività di web crawling al fine di prevenire conflitti con altre attività e/o sovraccaricare la rete del C.R. di Portici;
- **Configurazione**: sono state utilizzate le configurazioni ottimali dei parametri al fine di massimizzare la velocità di download;
- **Monitoraggio**: i test sono stati tenuti sotto controllo durante la loro esecuzione.

Primo Test

Test del 10/12/2015

- **Numero di agenti:** 16
- **Tempo di esecuzione:** ~8 h
- **Quantità di dati scaricati:** ~2,94 TB
- **Quantità di risorse scaricate:** 66.806.790 Pagine
- **Velocità di dati scaricati**: ~850 Mbps
- **Velocità di risorse scaricate:** ~2305 Pagine/Sec.

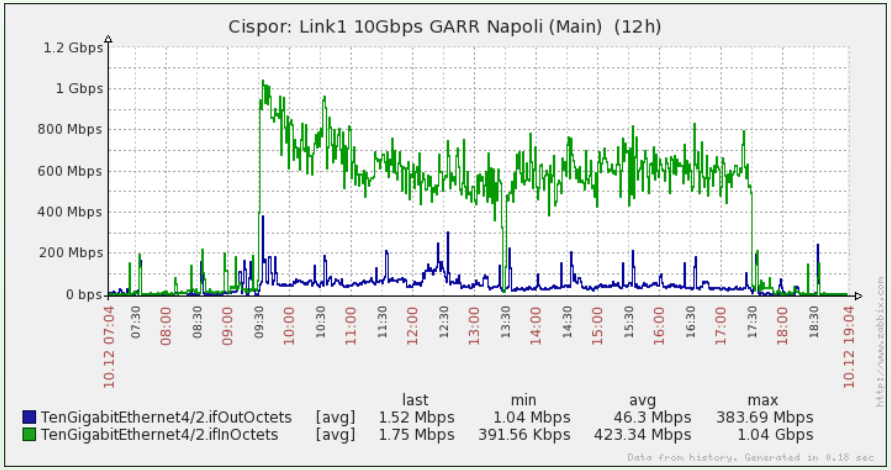
Traffico PoP Napoli-Portici



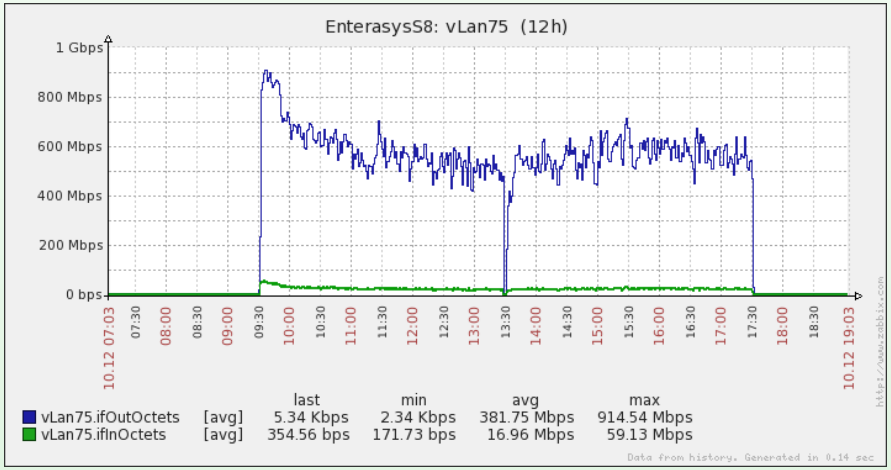
Primo Test

Test del 10/12/2015

Traffico C.R. Portici

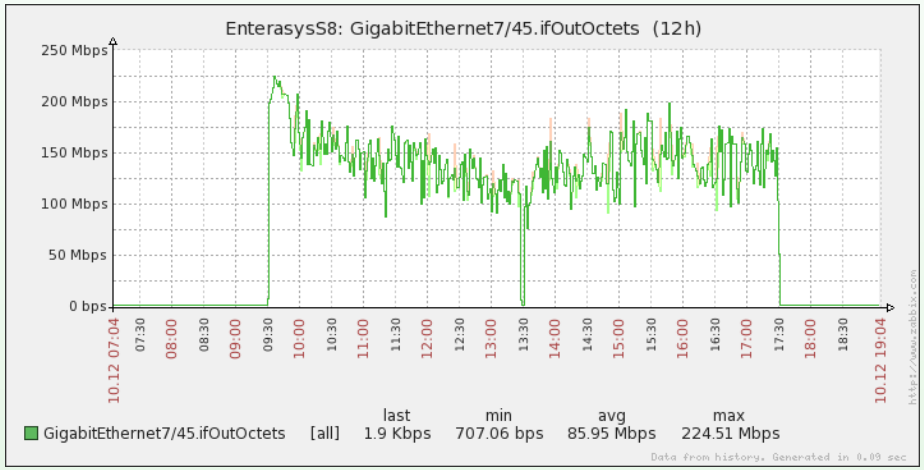


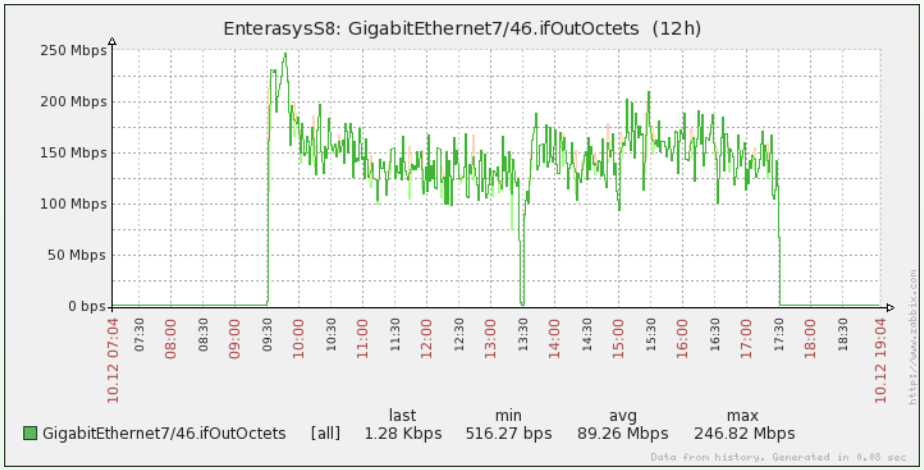
Traffico ENEA-GRID Portici

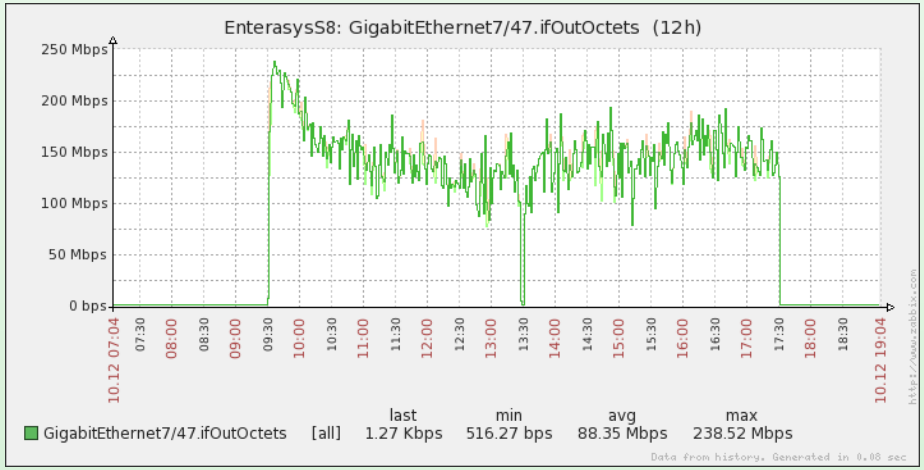


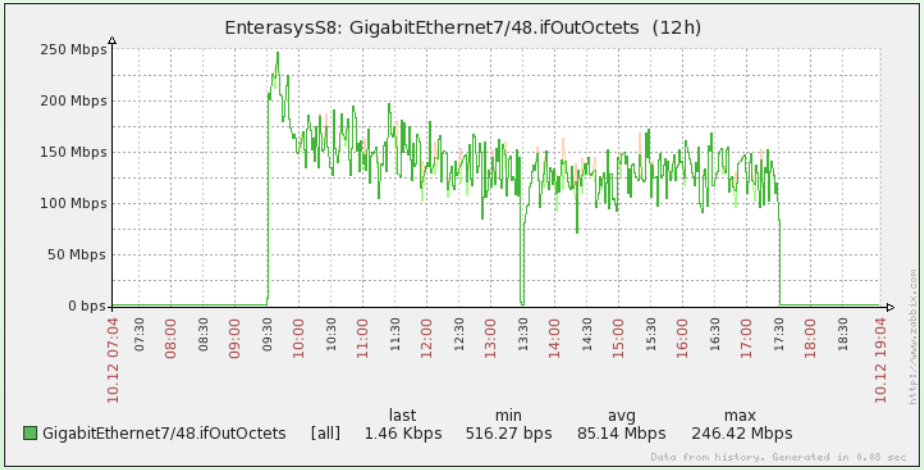
Primo Test

Traffico Nodi







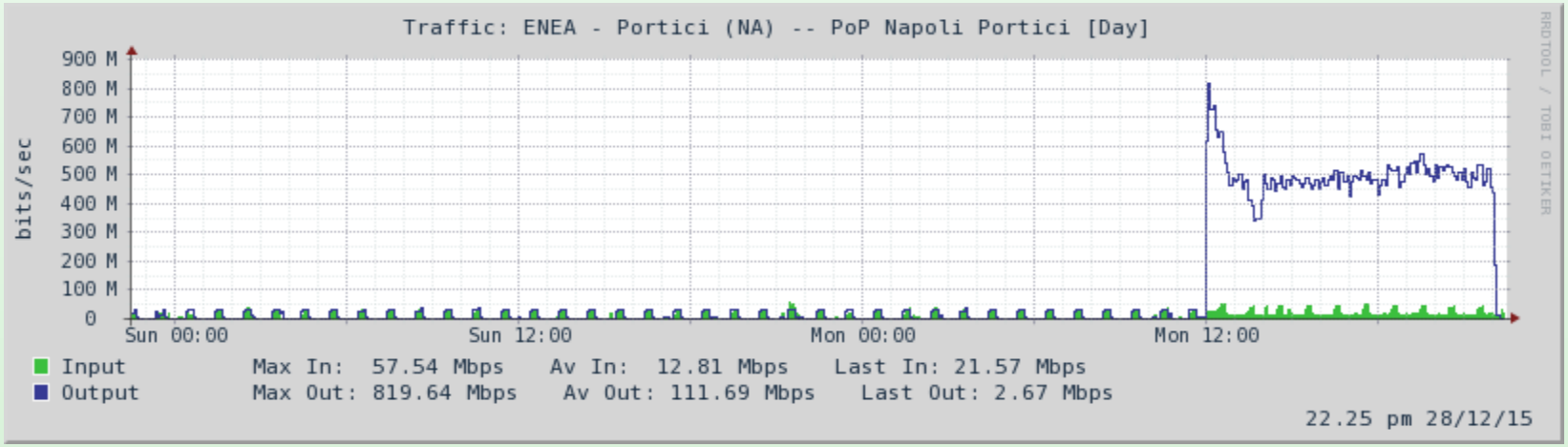


Secondo Test

Test del 28/12/2015

- **Numero di agenti:** 16
- **Tempo di esecuzione:** ~10 h
- **Quantità di dati scaricati:** ~3,23 TB
- **Quantità di risorse scaricate:** 71.667.304 Pagine
- **Velocità di dati scaricati**: ~740 Mbps
- **Velocità di risorse scaricate:** ~1959 Pagine/Sec.

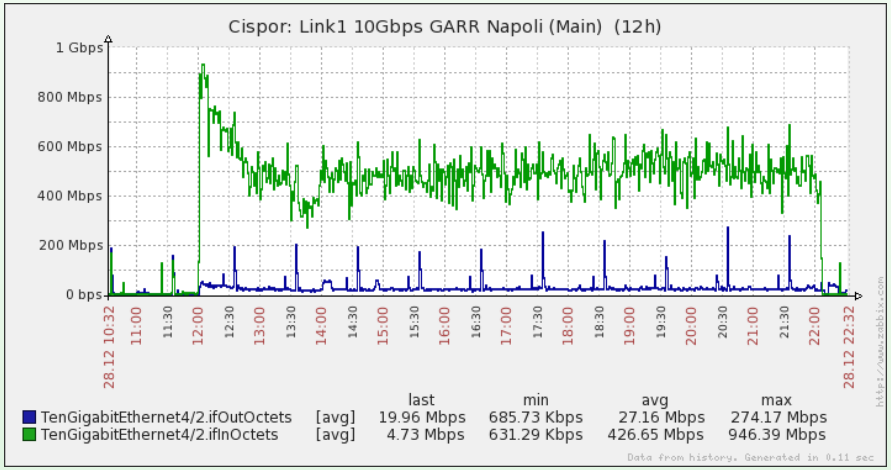
Traffico PoP Napoli-Portici



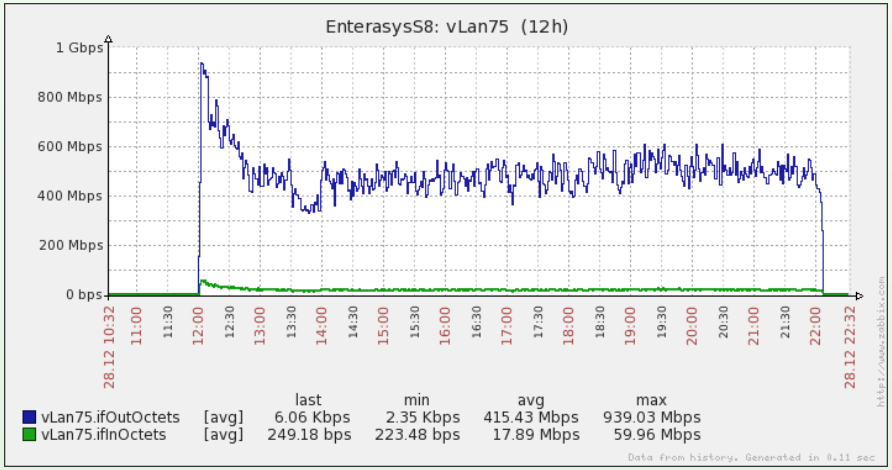
Secondo Test

Test del 28/12/2015

Traffico C.R. Portici



Traffico ENEA-GRID Portici

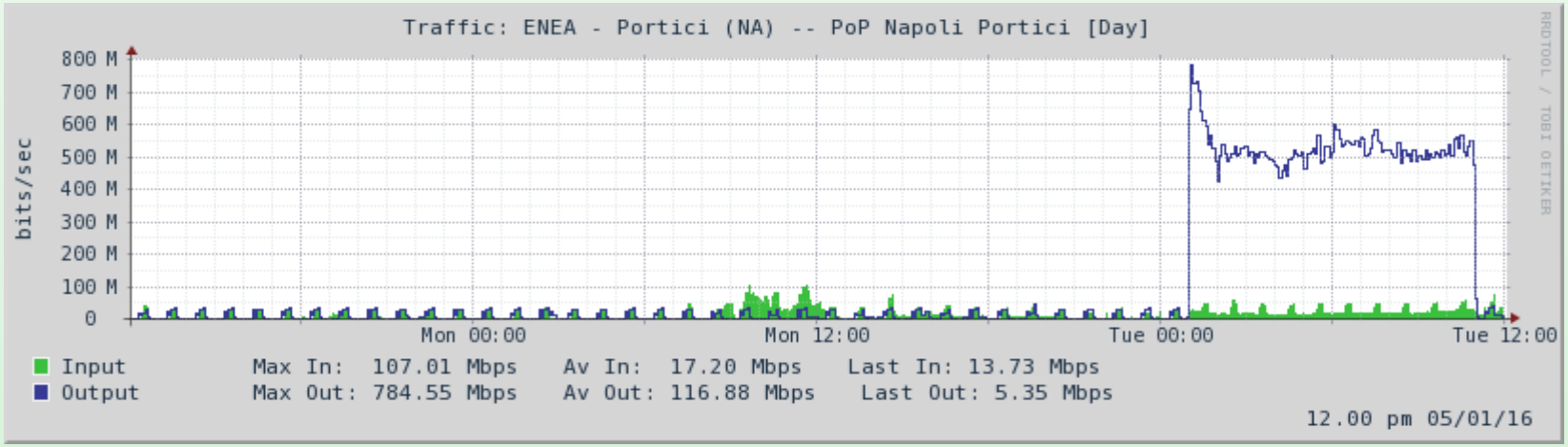


Terzo Test

Test del 05/01/2016

- **Numero di agenti:** 8
- **Tempo di esecuzione:** ~10 h
- **Quantità di dati scaricati:** ~3,27 TB
- **Quantità di risorse scaricate:** 75.587.287 Pagine
- **Velocità di dati scaricati**: ~756 Mbps
- **Velocità di risorse scaricate:** ~2084 Pagine/Sec.

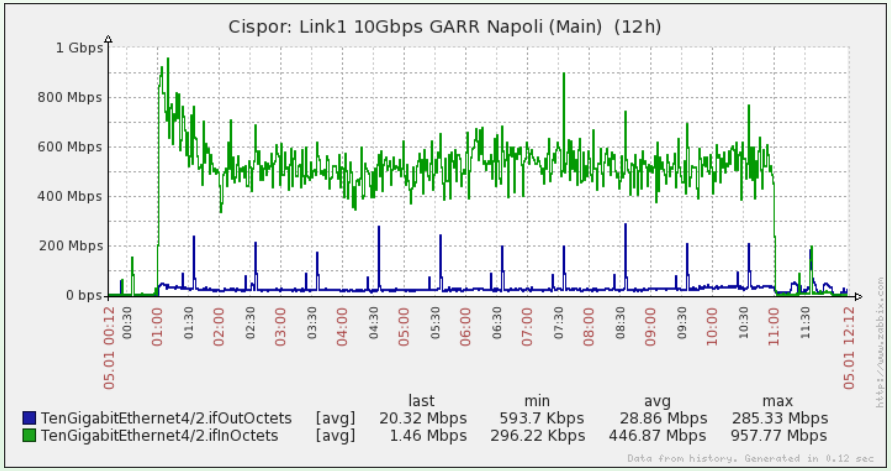
Traffico PoP Napoli-Portici



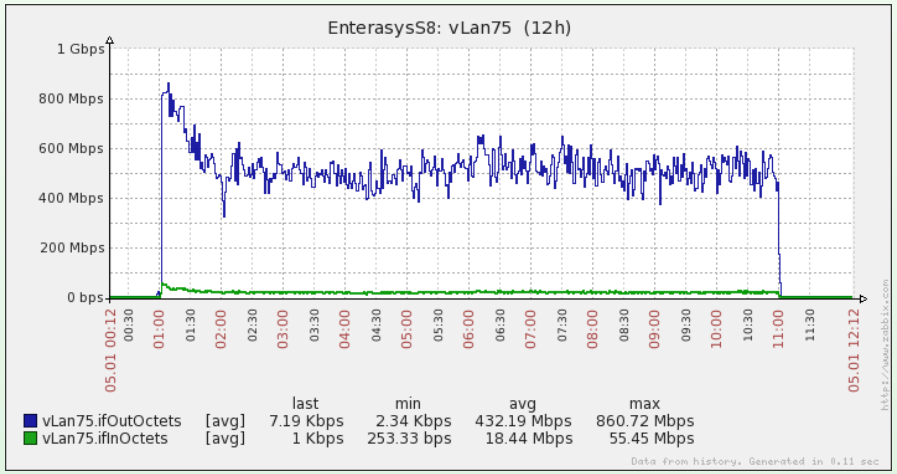
Terzo Test

Test del 05/01/2016

Traffico C.R. Portici



Traffico ENEA-GRID Portici



Analisi dei risultati

Osservazioni:

- Al momento, non sono state rilevate interferenze con la normale attività della rete del Centro di Ricerche ENEA di Portici;
- Durante le tre sessioni di crawling, la velocità di download ha avuto un picco iniziale e poi un attestamento su un valore costante più basso;
- Le migliori performance si sono avute con il primo test (~850Mbps);
- A parità di tempo di esecuzione (10h), la sessione con 8 agenti (~756Mbps) si è comportata meglio di quella con 16 agenti (~740Mbps).

Proposta per il rinnovo

I anno

- *Installazione e configurazione degli strumenti di web crawling.*

Rinnovo II anno: Proseguimento attività

- *Fruizione dei contenuti;*
- *Qualità dei dati;*
- *Analisi.*

Obiettivi:

- Investigazione sulla qualità e sulla natura dei dati scaricati;
- Studio e individuazione di opportuni strumenti per gestire, archiviare e visualizzare i dati (*big data*);
- Studio e selezione di tecniche per estrarre informazioni rilevanti e accurate dai dati provenienti dal web (*data mining*).

Proposta per il rinnovo

Attività previste:

- Crawling *semantico* e mirato su particolari topic;
- Snapshot di domini e identificazione di pattern;
- Definizione di tecniche e strumenti di indicizzazione e archiviazione dei dati;
- Utilizzo di tecniche di data mining per analisi avanzata dei dati;
- Creazione di un Laboratorio Virtuale sulla tematica del web crawling per integrare lo strumento e creare una community di lavoro collaborativo.

Fine

Grazie per l'attenzione.

giuseppe.santomauro@enea.it