

Sperimentazione del file system distribuito HDFS in ambiente grid

Giovanni Marzulli

INFN – Bari

Tutor: dott. Domenico Diacono



4° Borsisti Day – 13/09/2013

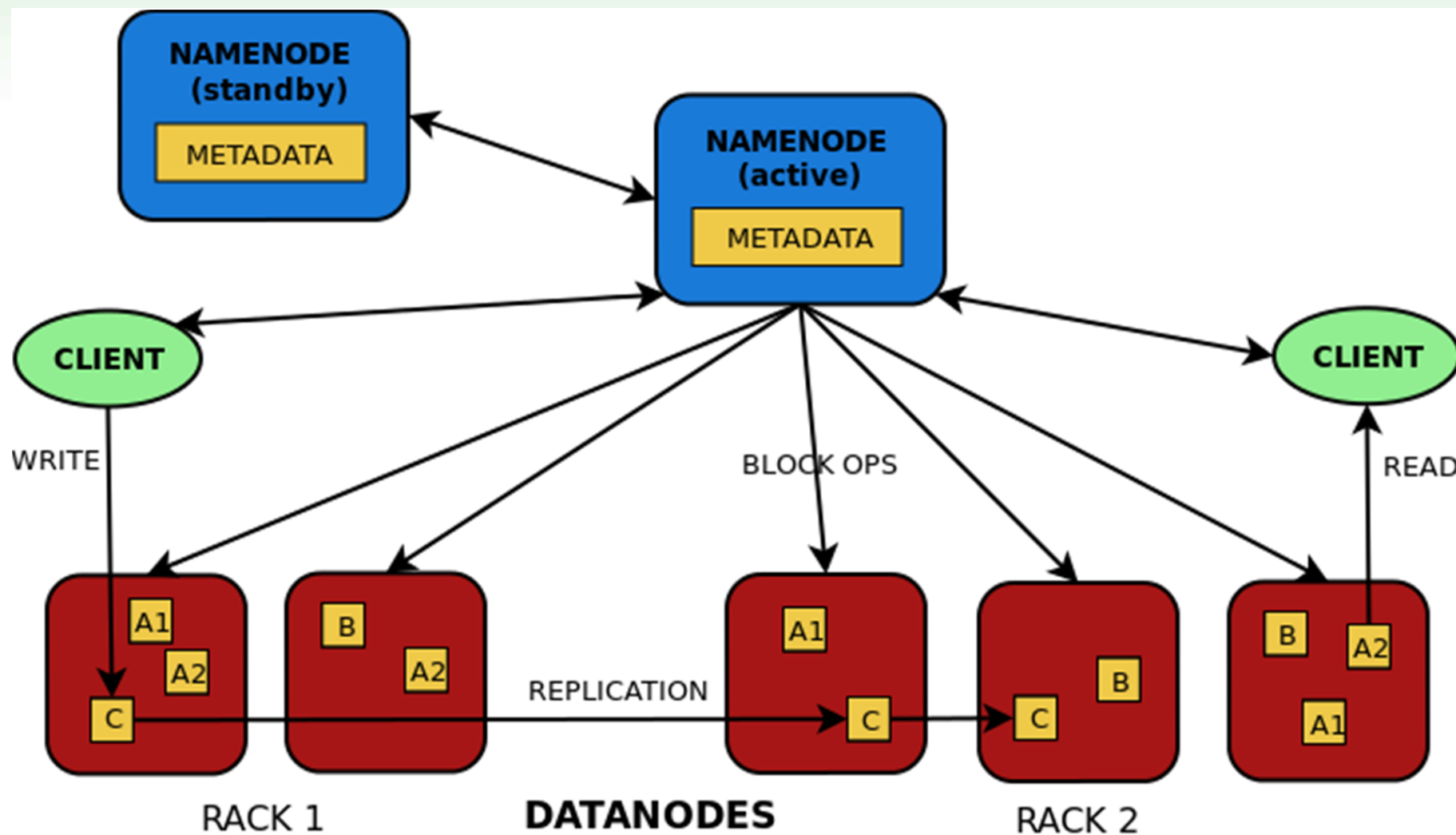


- Cosa è HDFS
- Attività svolta nel 2012
 - Test e sviluppo
- Attività svolta nel 2013
 - Netdisco
 - Monitoring avanzato
 - Test di performance
- Sviluppi futuri

Hadoop Distributed File System

- Open source
- Large dataset
- Fault tolerant
- Scalabile
- Commodity hardware
- Rack awareness

Cosa è HDFS: architettura



Attività svolta nel 2012



4° Borsisti Day – 13/09/2013

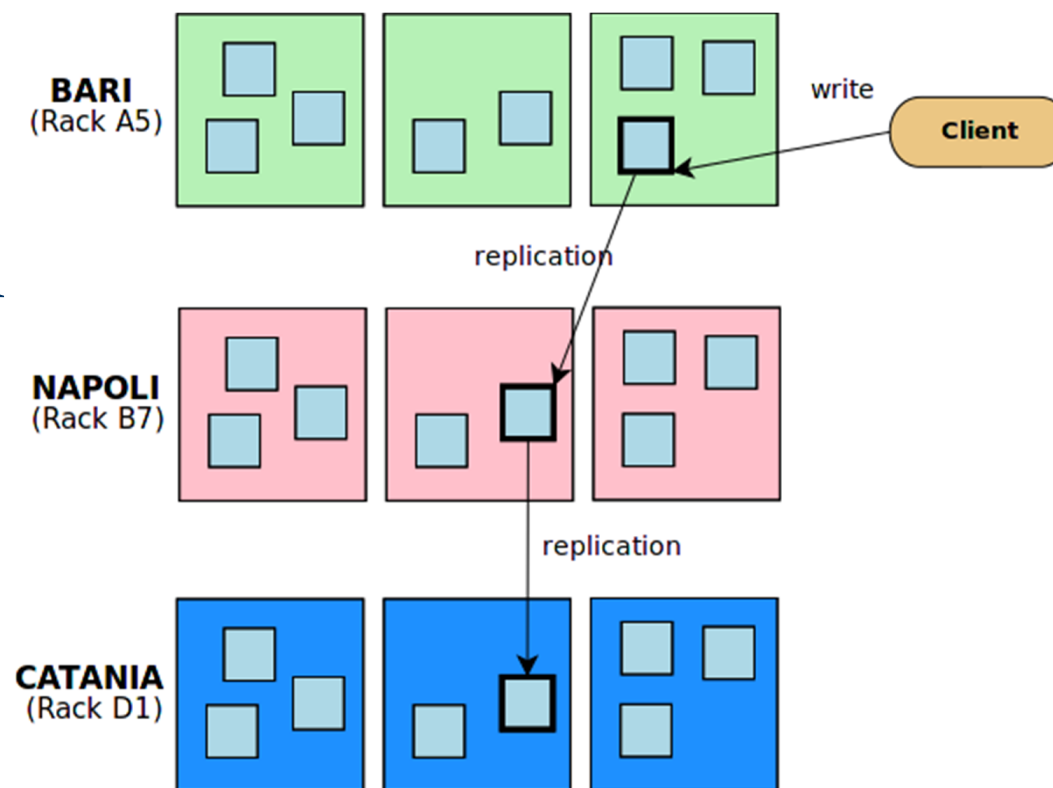


- Installazione e configurazione cluster di test in
 - LAN: infrastruttura locale INFN Bari
 - WAN: tra i siti INFN Bari e INFN Napoli
- Fault-tolerance test
 - simulazione di failure dei nodi (datanode e namenode)
 - failover e failback namenode
 - Ri-replica e ri-bilanciamento
 - under-replicated, over-replicated, mis-replicated blocks

- Autenticazione Kerberos
 - Nodi e utenti
 - Testata con il servizio Kerberos di Sezione
- Accesso ad HDFS
 - client Hadoop Java
 - Fuse-dfs

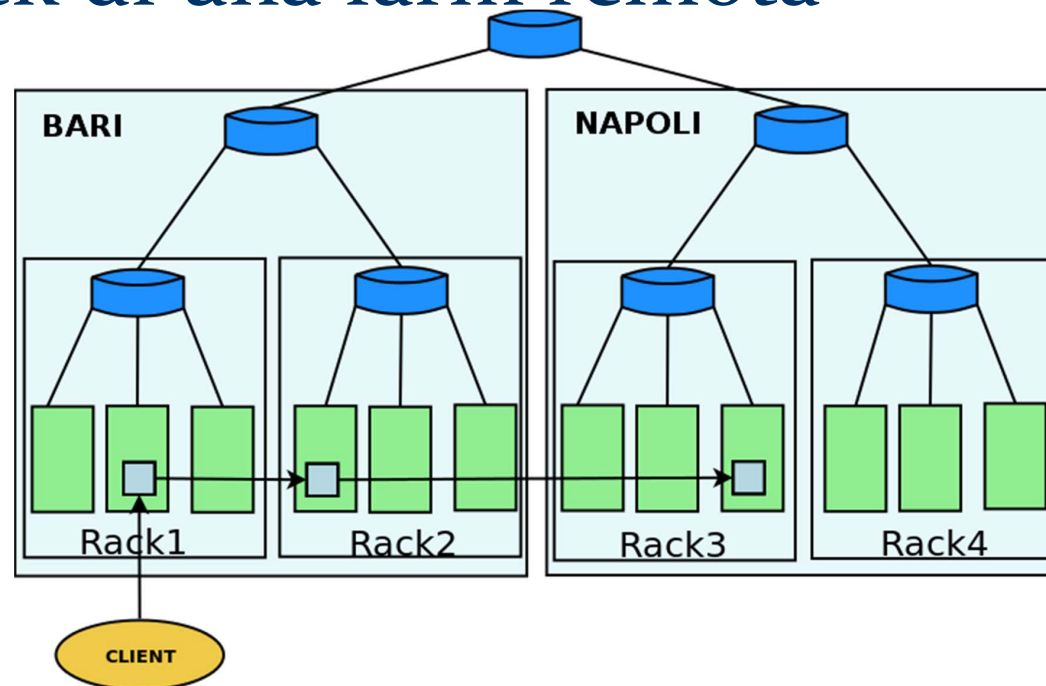
Sviluppo policy di replica: One replica policy

- 1 replica per rack
- Maggiore affidabilità
 - Tolleranza al fault di 2 racks (se fattore di replica 3)
- Utilizzata nel cluster HDFS in produzione a Bari



Sviluppo policy di replica: Hierarchical policy

- 2 repliche in rack differenti della farm locale
- 1 replica su un rack di una farm remota
- Tolleranza al fallimento di un'intera farm



Sviluppo di una procedura automatica di installazione e configurazione

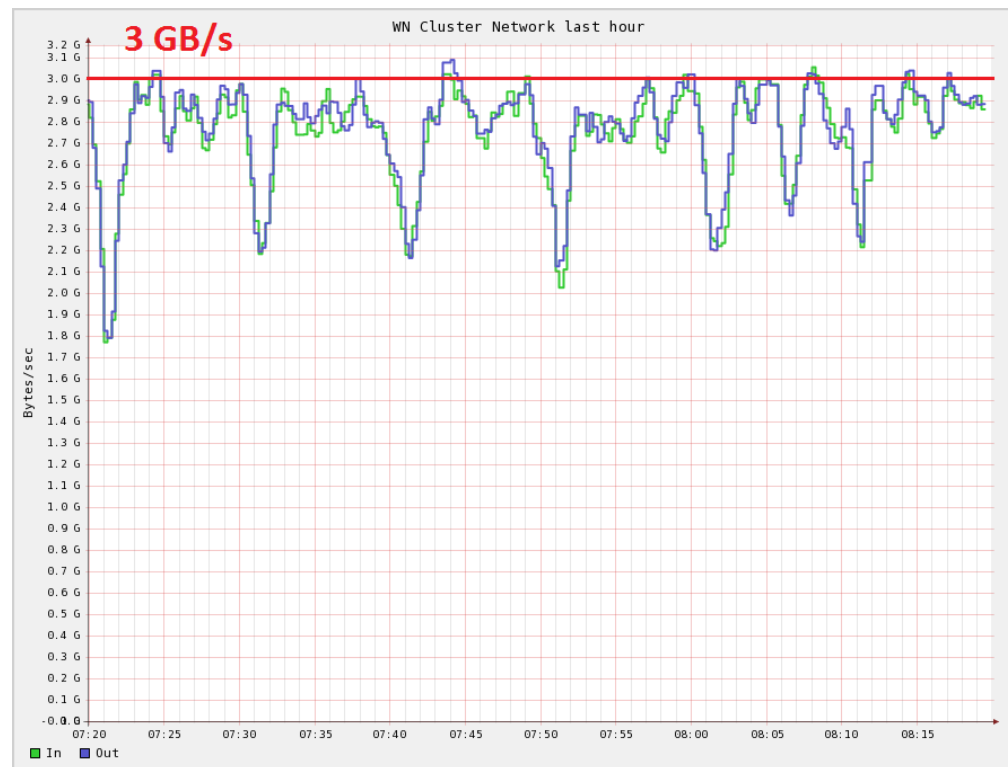
- Installazione del software
 - Packages repository
- Configurazione in base al tipo di nodo
 - formattazione, mount e assegnazione ad HDFS di dischi/partizioni inutilizzati
- Avvio (e riavvio in caso di fallimento) del processo

- Installazione e configurazione di Ganglia
- Custom monitoring, tramite analisi log in Bash
 - Locazioni dei blocchi dati
 - Recente cronologia
 - Blocchi danneggiati o persi
 - Operazioni sui blocchi
 - Memorizzati in un database MySQL

- Misurazioni di I/O con diverse configurazioni
 - Blocksize, fattore di replica
- Ambiente di test: centro di calcolo INFN Bari
 - Job di analisi in esecuzione
- Prestazioni (I/O sequenziale da singolo client):
 - Scrittura: 56,23 MB/s, Lettura: 61,56 MB/s

Test di performance: caso reale

- Prestazioni aggregate: 600 job concorrenti di Pamela in lettura via Fuse



Hadoop NameNode pccms61...

pccms61.ba.infn.it:50070/dfshealth.jsp Google

NameNode 'pccms61.ba.infn.it:9000' (active)

Started: Tue Nov 06 13:50:19 CET 2012
Version: 2.0.0-cdh4.1.1, 581959ba23e4af85afd8db98b7687662fe9c5f20
Compiled: Tue Oct 16 10:39:59 PDT 2012 by jenkins from Unknown
Upgrades: There are no upgrades in progress.
Cluster ID: CID-9b734b6d-3611-4eb7-ab5e-49419f75dc3a
Block Pool ID: BP-1130807058-212.189.205.51-1340275038748

[Browse the filesystem](#)
[NameNode Logs](#)

Cluster Summary

Security is OFF
370150 files and directories, 394102 blocks = 764252 total.
Heap Memory used 3.28 GB is 85% of Committed Heap Memory 3.84 GB. Max Heap Memory is 8.89 GB.
Non Heap Memory used 42.13 MB is 63% of Committed Non Heap Memory 66 MB. Max Non Heap Memory is 130 MB.

Configured Capacity	:	135.75 TB			
DFS Used	:	5.4 TB			
Non DFS Used	:	0 KB			
DFS Remaining	:	130.36 TB			
DFS Used%	:	3.98 %			
DFS Remaining%	:	96.02 %			
Block Pool Used	:	5.4 TB			
Block Pool Used%	:	3.98 %			
DataNodes usages	:	Min %	Median %	Max %	stdev %
	:	0.99 %	17.46 %	100 %	43.42 %
Live Nodes	:	227 (Decommissioned: 0)			
Dead Nodes	:	5 (Decommissioned: 0)			
Decommissioning Nodes	:	0			
Number of Under-Replicated Blocks	:	0			

NameNode Journal Status:
Current transaction ID: 4397350

Attività oggetto di due paper:

- «Using Hadoop File System and MapReduce in a Small/Medium Grid Site»
- «Testing and evaluating storage technology to build a distributed Tier1 for SuperB in Italy»

Attività svolta nel 2013



4° Borsisti Day – 13/09/2013



- Necessità di disporre di una lista sempre aggiornata della disposizione fisica dei datanode
 - macchina - armadio rack
 - switch etichettato con nome rack
- Installazione e configurazione di NetDisco
 - network discovery automatico via SNMP
 - scritto in Perl, DB PostgreSQL, Web UI

- Switch configurati solo a livello 2
- Realizzazione di uno script per la traduzione dei MAC address in IP address
 - Update nel DB
- Realizzazione di uno script per l'aggiornamento dinamico del file della topologia di HDFS
 - DB → file
 - Aggiornato giornalmente

Sviluppo di un sistema di monitoring avanzato

- Riprogettazione di un monitor avanzato per la gestione di un cluster complesso
 - gestione delle failures e dati di accounting
- Utilizzo delle API Java di Hadoop
 - Implementazione simile all'FSCK del client Hadoop



Applicazione Java



PostgreSQL



Web UI realizzata con
PHP framework Yii

Sviluppo di un sistema di monitoring avanzato: screenshot

HDFS Monitor - INFN Bari

[Home](#) [Nodes](#) [Storages](#) [Files](#) [Blocks](#)

[Home](#) » [Files](#)

Files

You may optionally enter a comparison operator (<, <=, >, >=, <> or =) at the beginning of each of your search values to specify how the comparison should be done.

[Advanced Search](#)

Operations

All

Healthy

Corrupt

Displaying 1-100 of 642457 results.

Uri	Status	File Size	Owner	Modification Time	Replication Factor
/pamela/deconvoluzione/deconvoluzione120517/plot_trk_m1_t0_e2_mi0/Histos080220-080317/cMeas.png	HEALTHY	0.01157283782959 MB	abruno	2012-05-17 10:55:23	2
/pamela/contaminations/step2/tof_without_calocut_0.15tol_eff.eps	HEALTHY	0.037181854248047 MB	abruno	2010-03-18 16:52:50	2
/pamela/deconvoluzione/deconvoluzione/KorrFlux_linear0/Histos081022-081118_cKorrFlux_linear_1.png	HEALTHY	0.023619651794434 MB	abruno	2012-01-05 12:22:08	2
/pamela/deconvoluzione/deconvoluzione/KorrFlux_notrk_1/Histos061017-061112_cKorrFlux_1.png	HEALTHY	0.020362854003906 MB	abruno	2012-02-14 11:31:34	2
/pamela/deconvoluzione/deconvoluzione/Korr_trk_1/Histos071228-080123_cKorr_1.png	HEALTHY	0.025127410888672 MB	abruno	2012-02-14 11:31:04	2
/pamela/deconvoluzione/deconvoluzione/Korr_trk_1/Histos090625-090721_cKorr_1.png	HEALTHY	0.025218963623047 MB	abruno	2012-02-14 11:31:05	2
/pamela/deconvoluzione/deconvoluzione_cutoff/plot_trk_old/single_plots/Histos061017-061112/Histos061017-061112_cKorrFluxEk_linear_34_1.png	HEALTHY	0.029746055603027 MB	abruno	2012-01-19 10:36:33	2
/pamela/deconvoluzione/deconvoluzione/data/flight/All_111125_trk_0_0/Histos090819-090914_cKorr.png	HEALTHY	0.025180816650391 MB	abruno	2012-02-13 22:54:28	2
/pamela/deconvoluzione/deconvoluzione/data/flight/All_111125_trk_0_0/Histos090915-091011.root	HEALTHY	0.039146423339844 MB	abruno	2012-02-13 22:52:40	2
/pamela/deconvoluzione/deconvoluzione/data/flight/All_111125_trk_0_0/Histos090915-091011_cKorr.png	HEALTHY	0.025362014770508 MB	abruno	2012-02-13 22:54:30	2

Sviluppo di un sistema di monitoring avanzato: screenshot

View File

#/mazziot/Fluka/Emissivity/EmissDataToTree_pN_dpmjet3_e)

Uri	/mazziot/Fluka/Emissivity/EmissDataToTree_pN_dpmjet3_extfluence_fl2011_2_17_tmp_R0.root
Status	HEALTHY
File Size	9526.9731035233 MB
Owner	mazziot
Modification Time	2013-01-20 17:21:05.487
Replication Factor	2
Located State	FALSE

Name	Block Size	Status	Corrupt
blk_3605258062379431038	64 MB	HEALTHY	TRUE
blk_-2221940351160802840	64 MB	HEALTHY	TRUE
blk_-9119052175070730714	64 MB	HEALTHY	TRUE
blk_-3046282723193494474	64 MB	HEALTHY	TRUE
blk_-6583566835666083256	64 MB	HEALTHY	TRUE
blk_7512560749691199422	64 MB	HEALTHY	TRUE

Sviluppo di un sistema di monitoring avanzato: screenshot

HDFS Monitor - INFN Bari

[Home](#) [Nodes](#) [Storages](#) [Files](#) [Blocks](#)

[Home](#) » [Blocks](#) » blk_3605258062379431038

View Block #blk_3605258062379431038

ID	57827
Name	blk_3605258062379431038
Block Size	64 MB
File	/maziot/Fluka/Emissivity/EmissDataToTree pN dpmjet3 extfluence fl2011 2 17 tmp R0.root
Status	HEALTHY
Corrupt	FALSE
Located State	FALSE

IP Address	Hostname	Storage ID	Status	Last Check
212.189.205.170	pccms141.ba.infn.it	DS-313582813-212.189.205.170-50010-1359549037552	LIVE	2013-06-14 05:15:00.245
90.147.66.202	alicegrid82.ba.infn.it	DS-1312972889-90.147.66.202-50010-1358963439078	LIVE	2013-06-14 05:15:00.23

- 3 mesi di produzione
 - Aumento della quantità di dati
- 3 milioni di inserimenti ad ogni scansione del file system per 20TB di dati
 - Tempo di esecuzione: oltre 2 ore per scansione
- Migrazione ad un No-SQL DBMS

Migrazione a No-SQL: scelta del DBMS

Cassandra

- Nodi peer
- Privilegia l'operazione di write
- Supporto nativo a CQL (SQL-like)
- Installazione immediata

HBase

- Basato su HDFS
- Privilegia l'operazione di read
- Nessun supporto ad un SQL-like
- Difficoltà di installazione

Upgrade del monitor

- Installazione di una istanza di Cassandra
 - 3 nodi su VMs con accesso a dischi reali
- Datastax Java API
 - Tempo di esecuzione: 45 minuti per scansione
- Hector Java API
 - Multi-threading, denormalizzazione
 - Tempo di esecuzione: 15 minuti per scansione
- Possibilità di ulteriori miglioramenti

Sviluppo componente sensore per Nagios

- Integrazione del monitor Hadoop sul sistema di monitoring centralizzato Nagios
 - Visione complessiva di tutti i servizi in produzione
- Sviluppo di una componente del sensore
 - Applicazione Java per l'interrogazione dello stato di HDFS tramite API
 - Produzione di whitelist e blacklist con indicazione del "last contact timestamp" per ogni datanode
 - Script per la determinazione dello stato del sensore

Test di performance: installazione

- Test realizzati su un cluster di 21 Macchine dedicate
 - 24 core, 80 GB RAM
 - 6 HD da 2TB
 - Link a 10Gbit/s wire speed SFP+
- Installazione software
 - OS: SLC 6
 - Hadoop: release del USCMS Tier 2 del Nebraska
 - Fuse-dfs stabile e performante
 - Ganglia

- I/O Singolo sequenziale
 - operazioni sequenziali eseguite su un solo nodo
- I/O Concorrente sequenziale
 - Operazioni concorrenti e sequenziali eseguite su tutti i nodi
- I/O Concorrente parallelo
 - operazioni concorrenti e parallele eseguite su tutti i nodi

Test di performance: fattori variabili

- HDFS
 - Fattore di replica: 2
 - Blocksize: 64MB, 128MB
- Fuse-dfs
 - rdbuffer: 64MB, 128 MB
 - Big_writes
- File: 4GB, 10GB
- Tool: dd, iofzone
 - bs: 1M, 4M

Test di performance: I/O singolo sequenziale

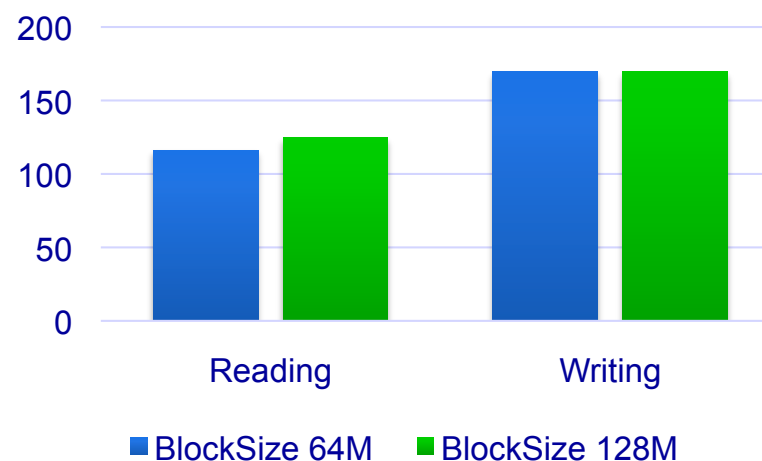
RdBuffer 128MB

MB/s	Block 64M	Block 128M
Reading	122,02	129,76
Writing	171,38	170,75



RdBuffer 64MB

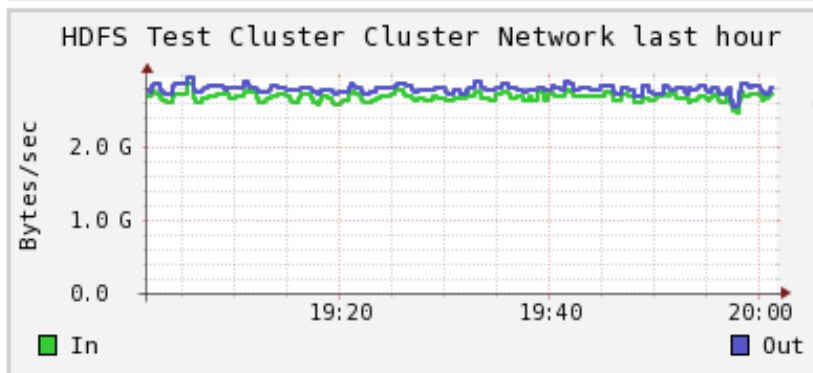
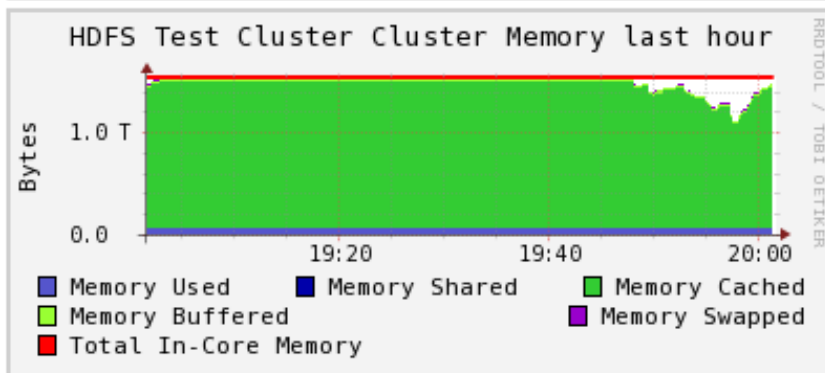
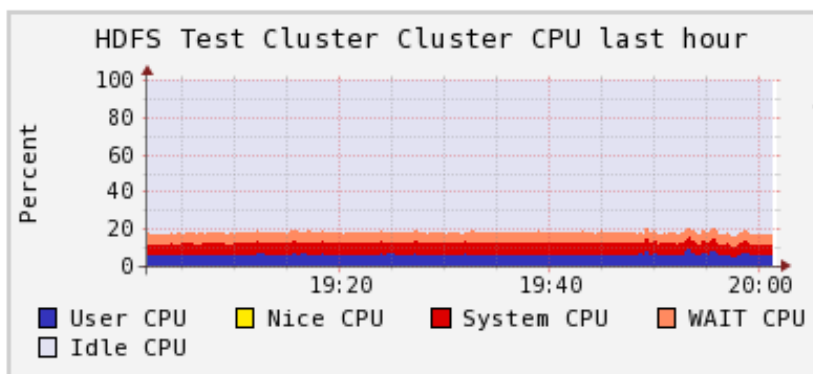
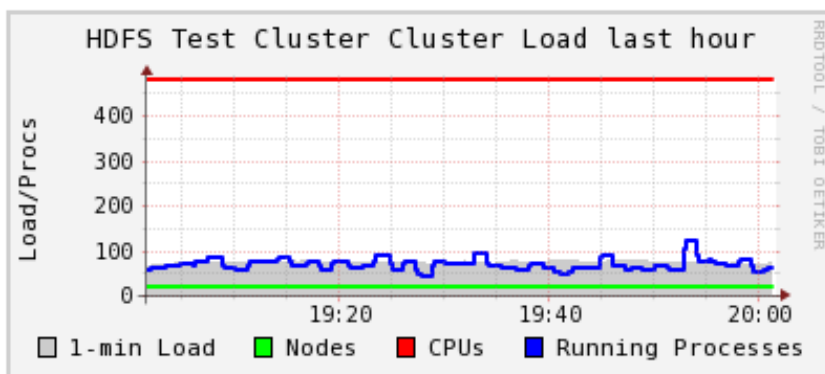
MB/s	Block 64M	Block 128M
Reading	116,18	124,56
Writing	169,60	169,41



Test di performance: I/O concorrente sequenziale

Writing (MB/s)

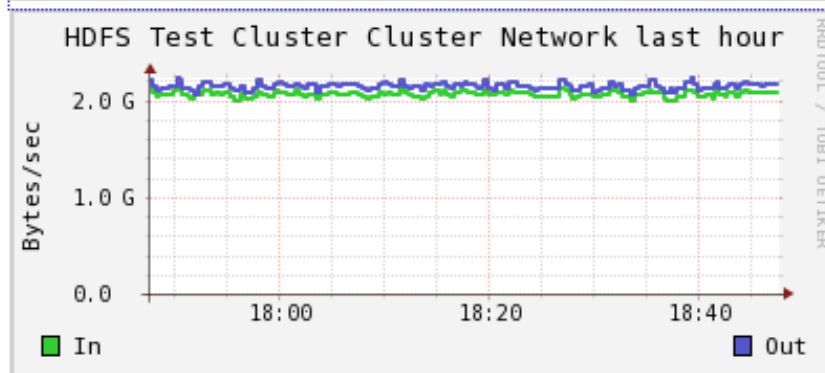
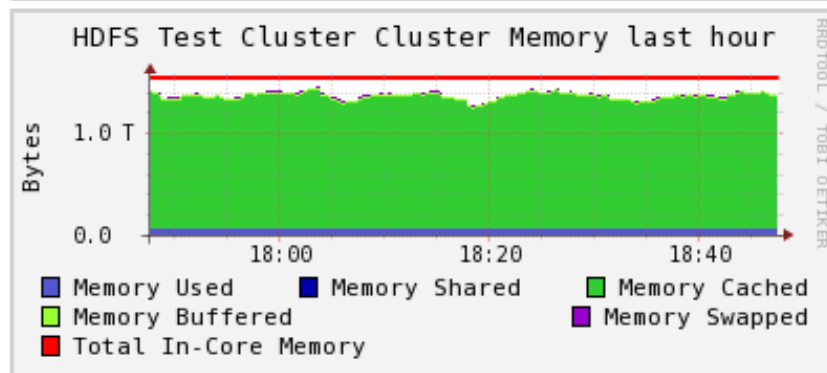
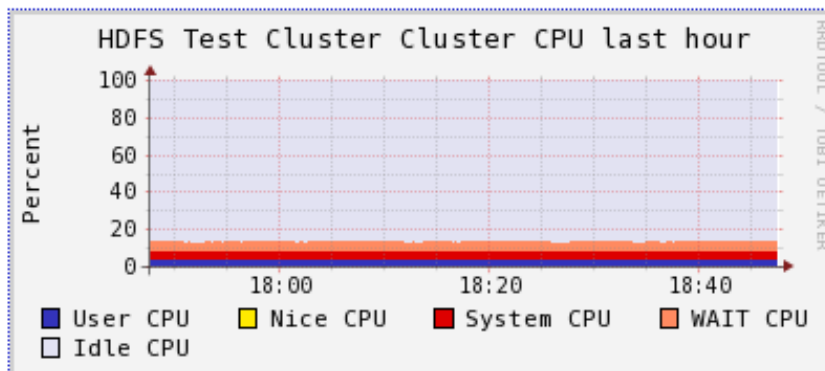
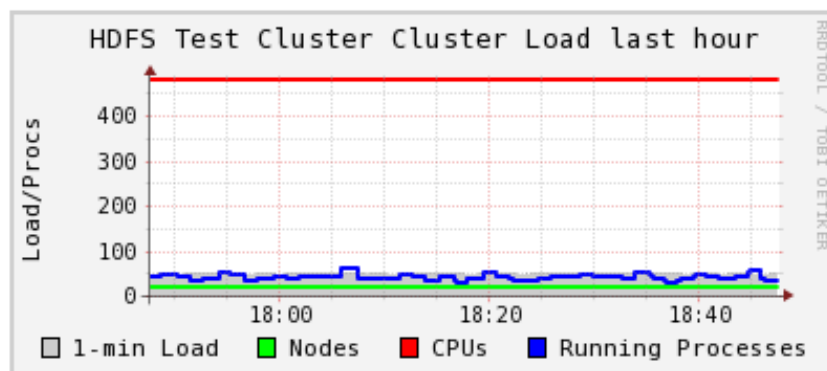
BlockSize / RdBuffer	64 MB	128 MB
64 MB	141,09	140,62
128 MB	142,22	143,51



Test di performance: I/O concorrente sequenziale

Reading (MB/s)

BlockSize / RdBuffer	64 MB	128 MB
64 MB	92,25	93,19
128 MB	110,12	113,53

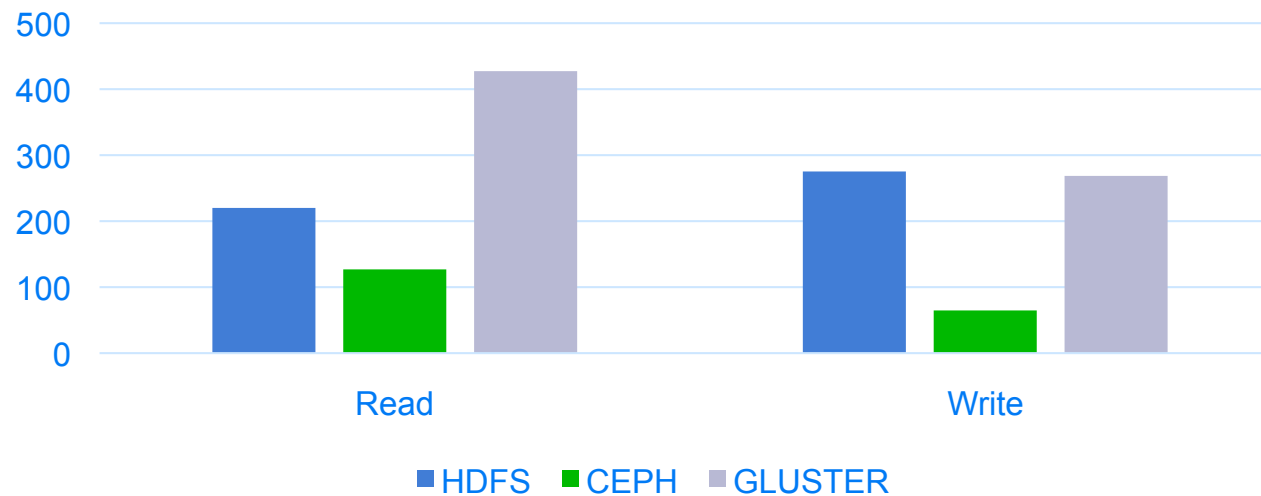


Test di performance: I/O concorrente parallelo

- Configurazione più performante:
 - Blocksize 128MB
 - rdbuffer 128MB
 - big_writes attivo
- Iozone con 24 threads
 - HDFS not fully POSIX-compliant
- 24 dd paralleli
 - Confronto con altri FS distribuiti

Confronto con altri file system distribuiti

MB/s	HDFS	CEPH	GLUSTER
Read	220,05	126,91	427,3
Write	275,27	64,71	268,57



Presentazione dei risultati al CHEP 2013

- Terminazione del processo di sviluppo del monitor HDFS
- Progetto SuperB recentemente soppresso
 - Test dell'infrastruttura tra 3 siti geograficamente distribuiti
 - Test di sistemi di sistemi di autenticazione avanzati con supporto alle federazioni
- Sottomissione delle policy ad Apache Hadoop
- Pubblicazione e rilascio del monitor

Grazie per l'attenzione



4° Borsisti Day – 13/09/2013

