

CONTRATTO DI RICERCA ENEA-CASPUR – CRESCO SPI.2

**ANALISI E REALIZZAZIONE DI TOOL INNOVATIVI A SUPPORTO DELLE
FUNZIONALITÀ GRID**

OR3 – OTTIMIZZAZIONE DELLE PRESTAZIONI DI AFS

Alessio Rocchi

Sommario

OR3 – OTTIMIZZAZIONE DELLE PRESTAZIONI DI AFS.....	1
1 - Introduzione.....	3
2 – Strumenti di fault monitoring.....	4
2.1 – Introduzione.....	4
2.2 – Stati d’allarme.....	4
2.2.1 – Volumi unattacheable.....	4
2.2.2 – Volumi unreleased.....	4
2.2.3 – Volumi off-line.....	5
2.2.4 – Volumi read-only con ID del parent non divulgato.....	5
2.2.5 – Volumi senza backup.....	5
2.2.6 – Volumi senza copia read-only.....	5
2.2.7 - Server Offline.....	5
2.2.8 – ID non regolari.....	5
2.2.9 – Indirizzi non risolvibili.....	6
2.2.10 – Volumi con parametri replicati.....	7
2.2.11 – Volumi Busy.....	7
2.2.12 – Volumi read-only su partizioni errate	7
2.2.13 – Volumi bloccati.....	7
2.2.14 – Volumi presenti solo nel volume database o nei file server.....	7
2.2.15 – Spazio in esaurimento sulle partizioni.....	7
2.2.16 - Partizioni overbooked.....	9
2.2.17 – Ritrasmissioni dei pacchetti spediti dai server.....	9
3 – Stima dei parametri ottimi.....	10
3.1 – Introduzione.....	10
3.2 – Il software	10
3.3 – Considerazioni e risultati sperimentali.....	11

1 - Introduzione

La linea di attività OR3, anch'essa orientata allo studio di OpenAFS, è stata caratterizzata dalla realizzazione di strumenti sia orientati al rilevamento di situazioni anomale o patologiche, sia deputati all'analisi inferenziale per la stima dei parametri ottimi di configurazione.

Ai fini infatti della ottimizzazione delle prestazioni del file system distribuito AFS risultano essere rilevanti vari elementi: (1) il controllo della corretta configurazione di funzionamento del sistema dal punto di vista dei server (2) il corretto stato operazione della infrastruttura di rete utilizzata (3) la scelta di parametri ottimali di funzionamento della parte client.

Nel seguito del documento verranno discusse in dettaglio l'analisi e gli strumenti utilizzati su questi tre fronti.

La sezione dedicata al "Fault Monitoring" in particolare si occupa dei primi due temi mentre la sezione dedicata alla "Stima dei parametri Ottimi" è dedicata alla ottimizzazione lato client.

2 – Strumenti di fault monitoring

2.1 – Introduzione

AMACA, tool descritto nella documentazione descrittiva dell'attività OR1, ingloba numerose procedure di analisi che rilevano e registrano, all'esecuzione di ogni snapshot, numerose situazioni di anormalità.

I moduli di reportistica di AMACA presentano questi dati all'attenzione dell'amministratore di sistema al momento in cui egli naviga, con il suo browser, verso la *home page* della web application. Essi lavorano in maniera adattativa: vengono visualizzati solo gli allarmi rilevati al termine della procedura di indicizzazione del file system, accompagnati dal numero di entità coinvolte nell'allarme.

La semplice visualizzazione degli stati d'allarme (che non comporti, cioè, una volontà o una necessità di effettuare un'analisi più approfondita) non richiede nemmeno l'inserimento delle credenziali di accesso.

Nel seguito del capitolo verranno elencati gli eventi critici monitorati dalle procedure interne di AMACA che permettono di identificare le problematiche relative a molte delle malconfigurazioni del sistema lato server.

All'ultimo punto verranno anche illustrate le capacità di autodiagnostica delle condizioni di rete.

2.2 – Stati d'allarme

AMACA è in grado di rilevare 18 stati d'allarme, che verranno descritti in dettaglio nei paragrafi successivi.

2.2.1 – Volumi *unattacheable*

Si tratta di un allarme grave: l'esperienza suggerisce che sia dovuto a problemi di storage o a notevoli problemi interni occorsi durante la creazione del volume.

Un volume rilevato come *unattacheable* è non responsivo e irraggiungibile.

2.2.2 – Volumi *unreleased*

Si tratta di un allarme di moderata gravità, e indica che un'operazione di creazione o update della copia read-write di un volume non è stata seguita da un opportuno *release*.

2.2.3 – Volumi off-line

Questo allarme rileva i volumi off-line.

2.2.4 – Volumi read-only con ID del *parent* non divulgato

Si tratta della segnalazione di una anomalia, fortunatamente di facile risoluzione.

AFS presenta questi volumi pubblicandone correttamente le informazioni che lo identificano come *read-only*, ma non espone in alcun modo le informazioni relative al volume read-write a cui esso afferisce.

Coinvolge volumi read-only mai rilasciati, ed è in genere sufficiente eseguire un *release* per rimediare al problema.

2.2.5 – Volumi senza backup

Si tratta di un allarme, potenzialmente grave, che rileva l'assenza, su volumi presenti in una lista compilata dall'amministratore, del corrispondente volume di backup. La sua creazione implica la sparizione del messaggio di allerta.

2.2.6 – Volumi senza copia read-only

Si tratta di un allarme che rileva l'assenza, su volumi presenti in una lista compilata dall'amministratore, della corrispondente copia read-only. La sua creazione implica la sparizione del messaggio di allerta.

2.2.7 - Server Offline

Questo avviso ha a che fare con la mancata raggiungibilità, da parte del Basic Overseer (BOS) server, di una o più macchine registrate.

2.2.8 – ID non regolari

Stando alla documentazione ufficiale, gli ID dei volumi read-write, read-only e backup devono seguire uno schema tale che, detto i l'ID del volume RW, i volumi RO e BK avranno ID pari ad $i+1$ e $i+2$. Non sempre questa numerazione è rispettata nella creazione dei volumi (per motivi di varia natura) e il sistema lo segnala in modo che l'amministratore possa provvedere a ripristinare una situazione regolare.

2.2.9 – Indirizzi non risolubili

Questo allarme rileva quei file server i cui indirizzi non sono risolubili tramite richieste DNS.

2.2.10 – Volumi con parametri replicati

Questo allarme è stato riscontrato una sola volta, durante l'indicizzazione della cella pubblica di CERN. Si tratta di segnalare una situazione patologica in cui lo stesso volume (con gli stessi parametri identificativi) compare riportato più volte nella lista ottenibile istanziando richieste *vos listvol*.

2.2.11 – Volumi Busy

Questo allarme identifica i volumi che, all'atto del *crawling*, risultavano in stato *busy*. Anche se una simile situazione potrebbe non risultare particolarmente critica (il flusso OOB di dati verso il volume potrebbe essere dovuto ad un'attività temporanea e assolutamente regolare), è corretto segnalare affinché l'amministratore provveda a controllarne la natura.

2.2.12 – Volumi read-only su partizioni errate

Per questioni di ottimizzazione nel trasferimento dei dati, se un volume ha una copia read-only sullo stesso server, essa deve trovarsi anche sulla stessa partizione.

Questa segnalazione è di rapida soluzione per mezzo dei comandi *vos move* e *vos release*.

2.2.13 – Volumi bloccati

Esistono alcune situazioni in cui un volume risulta in stato *locked*. AMACA le segnala in modo che l'amministratore possa provvedere (per mezzo del comando *vos unlock*).

2.2.14 – Volumi presenti solo nel volume database o nei file server

Si tratta, in realtà, di una coppia di allarmi distinti, volti a evidenziare il fatto che un set di volumi sia attualmente registrato solo nel VLDB (e allora occorre provvedere a rimuovere le relative *entries*, per il fatto che sembra non esistere alcuna corrispondenza "fisica" sui file server), o solo nei file server (ed è necessario indagare sulla motivazione di questo stato, per decidere una strategia di risoluzione opportuna).

2.2.15 – Spazio in esaurimento sulle partizioni

Questo allarme rivela che, sulle partizioni segnalate, esiste una disponibilità di spazio fisico in esaurimento.

La soglia è configurabile, e attualmente l'eccezione è sollevata per partizioni con spazio libero inferiore al 20% del totale.

2.2.16 - Partizioni *overbooked*

Questo allarme segnala una situazione potenzialmente a rischio: ai volumi della cella viene assegnata, in fase di creazione, una certa quota disco (che può restare anche inutilizzata, con la conseguenza che lo spazio effettivamente necessario alle entità registrate in tali volumi è notevolmente inferiore a quello teoricamente disponibile).

Quando la somma delle quote disco di tutti i volumi presenti su una partizione di un file server eccede la quota fisica associata alla partizione stessa, è opportuno provvedere ad un ridimensionamento dei volumi (o ad un loro spostamento su una partizione meno "impegnata", nel caso in cui lo *shrinking* non sia possibile).

La criticità dell'overbooking dipende fortemente dalla modalità d'uso del volume.

Nel caso dei volumi dedicati alla home degli utenti la quota è normalmente gestita in modo flessibile sulla base delle richieste e delle politiche generali di gestione dello spazio. Nel caso invece di volumi dedicati ad aree dati la disponibilità di spazio deve essere garantita al momento della esecuzione dei codici produttori di dati. In tal caso l'utilizzo di overbooking deve essere minimizzato.

2.2.17 – Ritrasmissioni dei pacchetti spediti dai server

Questo allarme identifica problemi di comunicazione tra file server (connettività limitata, *bottleneck* sui percorsi di routing, ...) che sono identificati da un elevata quantità di pacchetti rispediti (e quindi un alto valore del rapporto *resends/data*).

I dati sulla ritrasmissione di pacchetti vengono acquisiti attraverso l'utilizzo dell'utility **rxdebug**, uno dei componenti del pacchetto OpenAFS. Tale utility fornisce delle preziose statistiche riguardo alla comunicazione tra i cache manager e i processi del file server e permette una autodiagnostica della qualità della rete, elemento essenziale per il buon funzionamento di OpenAFS.

L'analisi storica di tale dato permette facilità l'identificazione di problematiche dovute a l'introduzione di nuovi componenti di rete malfunzionanti oppure causati da guasti di componenti esistenti.

3 – Stima dei parametri ottimi

3.1 – Introduzione

L'attività che conclude questo documento è stata relativa alla realizzazione di un tool che in modo automatico rilevasse le impostazioni ottime del cache manager del client di OpenAFS, in risposta a operazioni effettuate sul trasferimento di file aventi dimensione variabile.

Nel seguito verranno chiarite le modalità operative dello strumento software realizzato, e saranno riportati alcuni grafici esplicativi.

3.2 – Il software

Lo strumento è stato realizzato in modo che supportasse l'automatizzazione dei compiti di

- Variazione dei parametri di configurazione dei demoni OpenAFS
- Riavvio dei demoni
- Esecuzione delle operazioni di creazione file, con parametri decisi dal *tester*, tramite l'uso di **lmdd**
- Raccolta dei risultati

Riassumendo, siamo di fronte a due coppie di task, di cui la prima riguarda operazioni tipicamente amministrative (risolte scrivendo un semplice algoritmo di *patching* in linea), e la seconda ha a che fare con l'analisi quantitativa vera e propria.

In particolare, come già detto, abbiamo fatto uso del tool **lmdd** per la sua capacità di svolgere le operazioni di copia file come **dd (1)**, ma più adeguato ai nostri scopi per via del rapporto sui tempi e sul throughput rilevati durante lo svolgimento delle operazioni.

L'uso di **lmdd** ha previsto il trasferimento di alcuni semplici parametri (condivisi con **dd**):

- **of**, per settare la destinazione della copia
- **bs**, per settare la dimensione dei blocchi
- **fsync**, per il commit su disco.

Nel prossimo paragrafo verranno definite le condizioni sperimentali, e saranno riportati i risultati conseguiti.

3.3 – Considerazioni e risultati sperimentali

L'esecuzione del tool è stata schedulata sulla macchina cresco1x015 del cluster CRESCO di Portici, in modo tale da rilevare **tempo impiegato** e **throughput** nelle seguenti condizioni sperimentali:

- Creazione di 100 file da 25KB ciascuno
- Creazione di 1000 file da 25KB ciascuno
- Creazione di 10 file da 10MB ciascuno
- Creazione di 1 file da 100MB

Le precedenti operazioni sono state eseguite facendo variare la dimensione del chunksize della cache del client OpenAFS tra i valori 13 e 21 ¹.

La cache è gestita in modo da essere completamente mappata in RAM.

I risultati sperimentali sono illustrati in tabella 1 che mostra i tempi di esecuzione dei test ed in figura 1 che illustra l'andamento della velocità di scrittura.

	1000x2K	100x25K	10x10M	1x100M
13	0.48124	0.04442	1.82028	1.79992
14	0.4713	0.04404	1.92638	1.80734
15	0.47136	0.0449	1.83254	1.89474
16	0.46336	0.04446	1.76368	1.79472
17	0.46318	0.04454	1.73788	1.82244
18	0.46812	0.04446	1.83892	1.8524
19	0.4592	0.045	1.70196	1.80248
20	0.46104	0.04528	1.78604	1.7985
21	0.45736	0.04628	1.8036	1.75316

Figura 1 – tabella dei tempi di esecuzione dei test in funzione della dimensione del chunksize

Il diagramma dei throughput mostra il seguente andamento:

¹ Il chunksize è la dimensione dei cosiddetti V-files, che contengono una copia locale dei file richiesti al file system divisi in chunk. La dimensione di default è 64KBytes, ma può assumere valori espressi dalla relazione 2^{size} . Nel nostro caso, dunque, la *chunksize dimension* è stata fatta variare tra 8 KBytes e 2 MBytes.

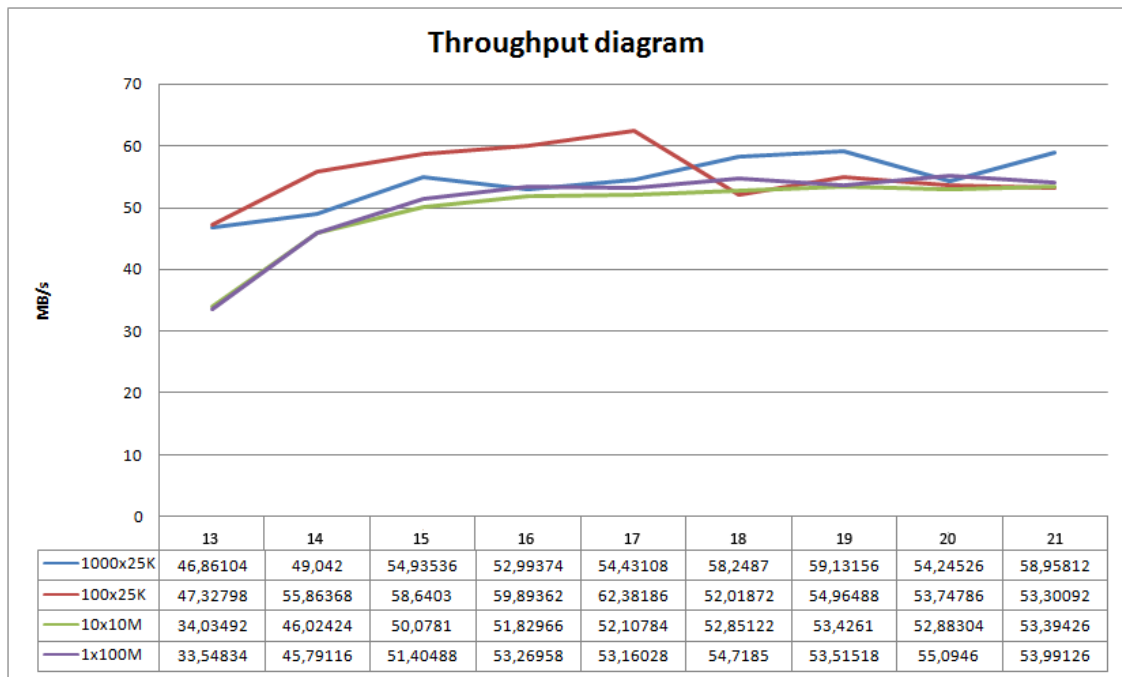


Figura 2 - Throughput in funzione della dimensione del chunksize

Il comportamento del throughput mostra che le prestazioni nel caso della configurazione scelta dipendono poco dalla dimensione di chunksize e che i valori di default, che a seconda della dimensione della cache, ricadono nell'intervallo 18-20 corrispondono a condizioni di funzionamento ottimale.

Nelle versioni più recenti di OpenAFS infatti la dimensione del chunksize viene scelta in modo automatico, a seconda della dimensione della cache disponibile.