Soluzione di Software as a Service (SaaS) per applicazioni scientifiche. Come sfruttare le risorse di calcolo e storage distribuito in grid tramite WorkFlow in modo semplice e trasparente.

Autori:

Giacinto DONVITO -- INFN-Bari
Pasquale NOTARANGELO -- INFN-Bari
Saverio VICARIO -- CNR - ITB
Bachir Balech -- CNR-IBBE

Abstract:

La disponibilità sempre più ampia di strumenti e tecnologie per produrre dati scientifici complessi e di dimensioni sempre più importanti sta rendendo sempre più necessario l'uso di complesse tecnologie di calcolo anche in contesti scientifici non abituati ad usare tali mezzi.
In questi contesti scientifici infatti i ricercatori sono abituati a lavorare con tool di alto livello per effettuare le proprie analisi, come ad esempio gestori di workflow, fra questi possiamo ricordare per esempio Taverna e LONI Pipeline.
Un gestore di workflow a differenza di quanto avviene in grid è costruito sulle interazioni sincrone, ed è costruito sull'ipotesi di sfuttare servizi affidabili e veloci.
Nel caso di una infrastruttura distribuita è invece necessario considerare che ogni singolo job potrebbe fallire o durare un tempo molto lungo. Inoltre, in alcuni casi una singola richiesta dell'utente può essere tradotta nell'esecuzione di centinaia o anche migliaia di job.
Per risolvere questi problemi è stato necessario costruire una interfaccia che facesse da gateway fra le richieste di un gestore di workflow e una grid di calcolo.
Tale gateway deve essere realizzato con una interfaccia che sia usabile da altri software e non direttamente dai ricercatori. La scelta è ovviamente caduta nell'implementare questa interfaccia come un WebService.
In questo lavoro verrà presentata l'attività di sviluppo portata avanti al fine di realizzare un layer di accesso alle risorse di calcolo distribuito in Grid, con particolare riferimento alla European Grid Infrastructure/Italian Grid Infrastructure. Tale layer si presenta come un WebService ed espone servizi con due protocolli: REpresentational State Transfer (REST) e Simple Object Access Protocol (SOAP).
Partendo da un tool di sottomissione automatica già usato per distribuire applicazioni di diverse scienze (bioinformatica, biomedicina, chimica computazionale, etc) sulla grid EGI/IGI è stato possibile supportare la sottomissione a diversi ambienti di calcolo: grid computing, farm locale, server dedicati. Con tale sistema è inoltre, possibile gestire la risottomissione dei job in caso di failure, il monitoring in tempo reale dello stato e di eventuali errori. Il tool è anche capace di

gestire la dipendenza fra job e questo è di notevole aiuto quando si devono gestire applicazioni particolarmente complesse.
La sottomissione dei job, ove necessario, è fatta attraverso l'uso di certificati Robot.
L'interfaccia a web service che è stata realizzata è in grado di supportare un numero ben definito di funzionalità:

- Sottomettere un job
- Controllare lo stato di un job
- Recuperare l'output di un job
- Sottomettere un numero arbitrario di job
- Controllare lo stato di tutti i job sottomessi nella stessa sottomissione e recuperarne lo stato
- Controllare lo stato di tutti i job di un determinato utente e recuperarne lo stato

Il job viene identificato con alcuni parametri che lo caratterizzano:

- Il software da usare
- alcuni parametri da fornire al software, fra cui i files di input
- lo stato
- un identificativo unico
- la localizzazione dell'output

Ogni applicativo dell'utente viene quindi "portato" su grid in modo da adattarne il modo di eseguirlo all'infrastruttura di calcolo distribuito.
L'utente deve solo occuparsi di chiedere l'esecuzione di una instanza di questo software con i parametri utili alla sua analisi, senza la necessità di gestire la complessità di eseguire una applicazione su internet. Infatti, non è necessario considerare problemi come:

- la risottomissione di job falliti
- la gestione del trasferimento dei file di input e di output
- la schedulazione in modo efficiente degli applicativi

Verrà, inoltre, descritta la soluzione implementata per la gestione dei dati basata su WebDaV che anche in questi contesti posso diventare di notevoli dimensioni.
Non è, infatti, raro in queste analisi avere input delle dimensioni di decine e anche centinaia di Gigabyte, oppure dover trasferire intere directory composte da centinaia o migliaia di files.
Tale soluzione di data transfer nel contesto di un workflow complesso, riesce inoltre, a minimizzare i traffici di dati fra il pc dell'utente e i servizi di calcolo favorendo lo scambio di dati interno alle infrastrutture di calcolo.

Verranno descritte, quindi, alcune applicazioni pratiche di questa soluzione di SaaS in cui un ricercatore è in grado di comporre workflow molto complessi mettendo insieme servizi classici di interrogazione sincrona, servizi basati sulle possibilità offerte dalle infrastrutture di calcolo distribuite, e servizi computazionalmente pesanti ma che richiedono una interazione veloce e che quindi risiedono su risorse dedicate ad alte performance.