ISTITUTO NAZIONALE DI FISICA NUCLEARE

Sezione di Pisa

INFN/CCR-09/05
23 Ottobre 2009

CCR-32/2008/P

RIFLESSIONI SULLA VIRTUALIZZAZIONE

Alberto Ciampa

INFN- Sezione di Pisa, Largo B. Pontecorvo, 3, I-56127 Pisa, Italy

Abstract

Proposta di un semplice contesto metodologico per la valutazione degli ambiti applicativi nei quali introdurre la virtualizzazione.

Vengono presentati e brevemente analizzati quattro scenari: *worker nodes on demand*, virtualizzazione per *high availability*, *farm on demand* e virtualizzazione per interattivo.

Si conclude con la presentazione della proposta per il coinvolgimento della Sezione di Pisa nel Gruppo di Virtualizzazione.

*Published by **SIS–Pubblicazioni***
Laboratori Nazionali di Frascati

1 ANALISI E VALUTAZIONE RISORSE

Come si ricava dall'insieme degli interventi presentati all'ultimo *workshop* CCR (2009) e dal riassunto della sessione sulla Virtualizzazione curato da Andrea Chierici, sull'argomento c'è molto interesse e sono in corso molte attività nell'ambito dell'intero Istituto. Le priorità del Gruppo Virtualizzazione dovrebbero quindi essere:

- il coordinamento delle varie attività in atto e future;
- l'armonizzazione delle azioni con un chiaro piano di obiettivi;
- la documentazione e la diffusione dei risultati.

La CCR può fungere allora non solo da promotore e catalizzatore del Gruppo sulla Virtualizzazione, ma innanzitutto come luogo per la definizione di strategie e obiettivi, raccolta delle esperienze, dei risultati e della relativa documentazione in materia, allo scopo di una sua diffusione e sfruttamento a livello di Istituto.

Segue una panoramica delle possibili attività di base:

- Valutazione delle soluzioni esistenti:
    - Le prestazioni in termini di efficienza di CPU non sembrano un importante elemento di merito (stiamo comunque parlando di oltre il 95% di prestazioni rispetto ai sistemi nativi)
    - Soprattutto tenendo conto delle applicazioni *target* è importante una attenta valutazione delle modalità e prestazioni di I/O. In alcuni ambiti di applicazione la virtualizzazione comporta complessi problemi di I/O e/o di gestione delle periferiche, in altri ambiti invece è necessaria un'elevata efficienza. In altri casi al contrario questi aspetti sono meno critici
    - Una analisi attenta va riservata alle differenti scelte architetturali e ai relativi scenari implementativi, in particolare per quel che riguarda il loro impatto sulla gestione sistemistica che ne consegue.
    - Si possono già individuare potenziali esigenze per una futura migrazione da una tecnologia di virtualizzazione ad un'altra. In generale per le diverse soluzioni andrebbe anche valutata la possibilità/facilità di migrazione, sia in ingresso sia in uscita.
- Valutazione su basi certe e contestualizzate dell'eventuale valore aggiunto di soluzioni proprietarie (o comunque a licenza) rispetto a quelle "open source" in termini di caratteristiche, prestazioni e servizi (in prospettiva stiamo pensando di far diventare questi *software* elementi della catena di produzione).
- Indagine sugli strumenti esistenti di gestione di una infrastruttura di virtualizzazione in relazione ai diversi ambiti nei quali si intende introdurla. Per esempio riflettere sulle esigenze e possibilità di "Fabric Virtualization": abbiamo tecnologie per la gestione di LAN complesse di macchine fisiche, ma per analoghe architetture LAN di *virtual machine* (VM)? (Pensiamo a SAN, *file system* distribuiti e paralleli, reti veloci ed in generale LAN *fabric* eterogenei).

A questo proposito riporto i punti salienti del *summary* della sessione sulla virtualizzazione tenutasi all'ultimo Hepix di Umeå (spring 2009) come riportati dal *chairman*,

Thomas Finnern (DESY):

- *Virtualization is more complex than standard technology*
- *The technology of host virtualization has reached production quality. Management systems - either commercial or open source – are available providing high level operating and scaling support*
- *It would be helpful to have a tight integration of virtualization into batch computing in order to ease middleware distribution and dynamic support of different operating systems*
- *Using imported, static or other miscellaneous operation system images e.g. for running in a cloud or grid might easily break privacy and security standards. There is currently no protecting sandbox available*
- *At first glance cloud computing appears to be a new promising technology. Until now it lacks of standard interfaces, cheap data access, security and other essential features, but we (and our users) will see a fast evolving (commercial) market.*

Un punto metodologico di partenza per l'impostazione delle attività potrebbe essere l'identificazione delle caratteristiche di base alle quali si guarda valutando l'introduzione della virtualizzazione in uno specifico ambito applicativo (i "driver"). In prima analisi è possibile introdurre due parametri, visti come assi ortogonali che individuano un piano sul quale collocare i diversi ambiti applicativi oggetto dello studio:

- Obiettivo di merito cercato:
  - Flessibilità: *server* con configurazione "a la carte" (caratteristica dell'"on demand").
  - Efficienza: ottimizzazione dell'uso delle risorse. Più VM sullo stesso *host* e soprattutto riduzione di *host* allocati ma non utilizzati.
  - Robustezza: *High Availability*, virtualizzazione come evoluzione della ridondanza per sistemi critici.
- Contesto di utenza:
  - Stabilità delle richieste (dinamica): la flessibilità non è gratis, introduciamola dove conviene.
  - Controllabilità degli utenti, *security*: le VM non possono avere il livello di sicurezza degli *host* nativi. Sappiamo chi utilizza le VM e (tendenzialmente) come?

Considerando le attività di base precedentemente enunciate si individuano quattro ambiti applicativi nei quali si può valutare come promettente l'introduzione di tecnologie di virtualizzazione.

Nella figura che segue (FIG.1) ho ipoteticamente posizionato gli ambiti applicativi, di seguito analizzati, sul piano dei *driver*:

- il posizionamento sulle "x", "Obiettivo", indica l'insieme di obiettivi *target* dello specifico ambito applicativo (obiettivo che si vuole raggiungere);

- il posizionamento sulle “y”, “Contesto”, indica la misura dell’impatto delle due caratteristiche del contesto (questioni e potenziali difficoltà delle quali tener conto alla partenza).

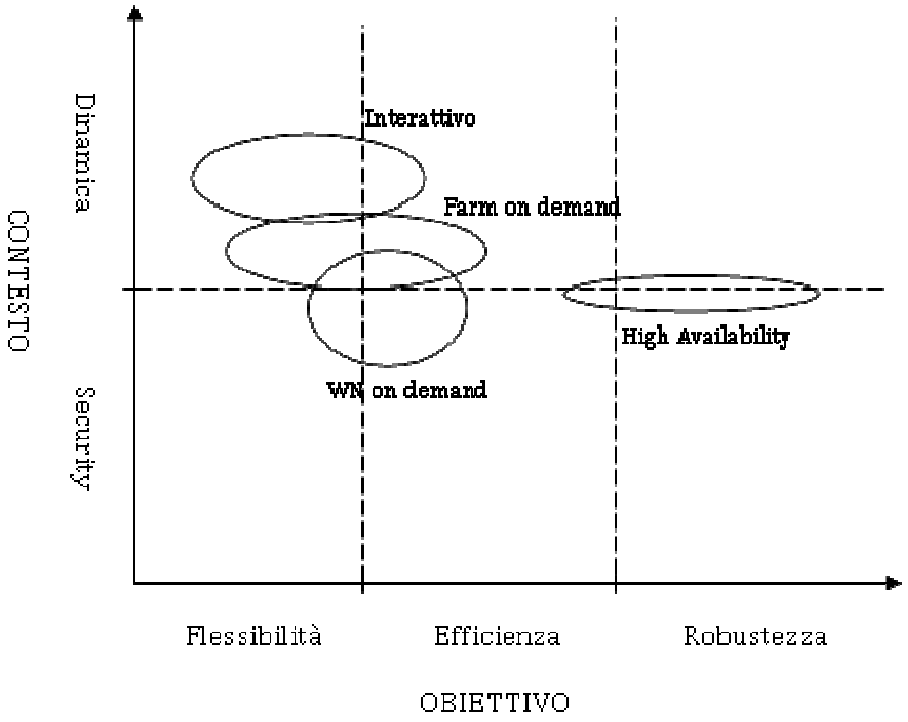

FIG. 1: Ambiti Applicativi e Driver.

2 WORKER NODES ON DEMAND

È il contesto sul quale si concentrano le maggiori attenzioni, probabilmente data la quantità di *server* utilizzati in guisa di *worker node* (WN).

Tentando una prima e sommaria analisi a partire dai “driver” riportati, proporrei le seguenti osservazioni:

- Esigenza di flessibilità. Senza considerarne i costi, la flessibilità è sempre desiderabile; parlando di WN andrebbe valutata l'effettiva necessità sulla base delle diverse tipologie di siti (grandi siti che hanno produzioni da molte VO e con pesi relativi mutevoli, siti più piccoli quasi mono-VO o siti che hanno molta produzione da VO “opportunistiche”).
- Ricerca di efficienza. In molte produzioni maggiori, specialmente di LCG in ambito INFN-GRID, la questione dell’efficienza di utilizzo della CPU rappresenta un punto di potenziale miglioramento. Andrebbe valutato se la virtualizzazione, per esempio con possibilità di eseguire più di un *job* per *core* contemporaneamente, rappresenti una

strada che offre vantaggi (essenzialmente migliori possibilità di gestione).

- Robustezza. Forse non è il motivo principale per cui si pensa di virtualizzare i wn.

Invece per quanto riguarda le caratteristiche del contesto di utenza:

- Stabilità delle richieste (dinamica). Andrebbe valutata come metro da usare per l'esigenza di flessibilità. La dinamica delle richieste delle maggiori VO richiede l'alto livello di flessibilità offerto dalla virtualizzazione (giustificandone i costi di vario tipo) o sarebbero sufficienti, almeno per ora, strategie più conservative? Alla domanda precedente si può dare una risposta generale o conviene differenziare la situazione in due o tre gruppi (per esempio in base alla classificazione dei siti menzionata in precedenza)?
- Controllabilità degli utenti (*security*). In termini generali un WN è tra i *server* di cui con maggiore difficoltà è possibile prevedere chi ne farà uso e come.

Ci sono due Sezioni (Perugia e Bari) più il CNAF che stanno lavorando in questa direzione, incontrando problemi comuni pur seguendo approcci differenti. Si è dichiarato interessato ad esplorare questa strada anche LNGS. È quindi già a lavoro una massa critica ben sufficiente a definire una attività specifica all'interno del Gruppo.

3 VIRTUALIZZAZIONE PER *HIGH AVAILABILITY*

In una sala calcolo tipica di una Sezione dell'INFN ci sono diversi sistemi che erogano Servizi a vario livello di criticità: sia Servizi generali verso la comunità di utenti che afferisce alla Sezione (gestione account, autenticazione e autorizzazione, DNS, DHCP, web ecc.) sia elementi portanti della struttura GRID (CE, SE, SRM, gestore code, ui ecc.).

In questo caso la virtualizzazione può essere interessante. Analizziamola sommariamente secondo i *driver* individuati:

- Obiettivi.
    - Flessibilità: non un obiettivo primario. I servizi erogati hanno caratteristiche stabili e la necessità di migrazione da un *host* all'altro è dettata dalla gestione di situazioni di emergenza.
    - Possibilità di maggiore efficienza: alcuni servizi sono scarsamente esigenti in termini di risorse, ma può essere rischioso collocarli tutti su un singolo sistema. Con la virtualizzazione si possono mantenere *server* "logici" dedicati a ciascun servizio con la possibilità di una dislocazione ottimizzata e dinamica su diversi *server* fisici.
    - Necessità di robustezza: si usa la virtualizzazione come alternativa evoluta della ridondanza di sistemi, è l'obiettivo principale di questo ambito applicativo.
- Problematiche di contesto:
    - Dinamica e *Security*. Non si hanno difficoltà dovute a richieste dinamiche in configurazione né relativamente alla controllabilità degli utenti o delle applicazioni. L'impatto sulla gestione sistemistica potrebbe risultare

largamente compensato dai benefici in robustezza ed efficienza.

A Pisa usiamo il sito GRID della Scuola Normale Superiore come *testbed* per diverse attività di sviluppo e tutti i servizi GRID girano su macchine virtuali. Sono stati eseguiti anche estensivi test di *crash* (e generale *unavailability* di servizi) con *recovery* totale e automatica [1), 2), 3)].

Nella direzione di virtualizzare dei Servizi critici si stanno già muovendo diverse Sezioni (5 DNS e auth*, 6 web, 7 altri servizi): con il necessario coordinamento una attività in tal senso è facilmente definibile.

4 FARM ON DEMAND

In molte Sezioni dell'INFN sono presenti "Farm di Esperimento", paradigma di calcolo scientifico prevalente nell'Istituto prima dell'avvento di GRID. A Pisa è in corso un'analisi sul profilo di utilizzo di tali *farm* (ne abbiamo 9) con risultati preliminari che indicano una percentuale di saturazione relativamente bassa (intorno al 33%). Da qui è nata l'idea di pensare alla realizzazione di una "meta farm" di Sezione capace di istanziare, mediante virtualizzazione, le singole *farm* esistenti (o nuove) a seconda delle richieste. In tal caso si avrebbe il seguente scenario:

- Obiettivi:
  - Flessibilità. Considerando il "catalogo delle macchine" inteso come la definizione del *server* "standard" per ciascuna *farm* (una sorta di "menu delle farm a la carte"), mediante virtualizzazione ciascuno dei *server* fisici della "meta farm" può ospitare *on demand* un *server* virtuale di ognuna delle *farm* "a catalogo".
  - Efficienza, ottimizzazione. Risulta evidente come questa possa rappresentare una via pratica per aumentare la percentuale media di utilizzo dei *server* originariamente dedicati alle *farm* di esperimento.
  - Robustezza. Sicuramente saremmo in grado di aumentare la *availability* media dei componenti delle *farm* attive, ma non è l'obiettivo trainante in questo ambito.
- Problematiche di contesto:
  - Dinamica e *Security*, cioè stabilità delle richieste e controllabilità degli utenti. Non dovrebbero presentarsi problemi su questi fronti, soprattutto facendo un confronto con altri contesti: lo "spazio" delle configurazioni che possono essere richieste è tenuto sotto controllo e la dinamica specifica delle richieste non dovrebbe essere eccessiva (generalmente non si chiede una *farm* per poche ore). Ogni *farm* istanziata farebbe, inoltre, riferimento ad una comunità di utenti chiusa e ben definita.

In questa ottica una promettente attività del Gruppo potrebbe essere, identificate un paio di Sezioni con alcune *farm* di esperimento attive, iniziare gradualmente a realizzare "farm virtuali" da dare in uso alle relative comunità di utenti. A Pisa stiamo impostando una attività

in tal senso mettendoci in contatto con i titolari delle *farm* meno usate (utilizzo medio su anno <10%) e prospettando loro la possibilità della virtualizzazione. Un punto tecnico da affrontare riguarderà lo spazio *storage* delle *farm* virtualizzate, da organizzare (ri-organizzare) per poter essere accessibile ad ogni istanziazione.

5 VIRTUALIZZAZIONE PER INTERATTIVO

L'ambito applicativo discusso nel paragrafo precedente può essere collocato sulla "frontiera" di GRID: abbiamo gruppi di utenti che, per vari motivi, non vogliono/possono usare GRID per la loro attività *batch*, ma le loro richieste non sarebbero concettualmente incompatibili con tale paradigma (infatti chiedono una *farm* e la usano spesso in *batch*). Uno dei risultati dell'analisi sul Calcolo Scientifico a Pisa è stato constatare (misura) di quanto l'uso in GRID sia meno costoso di un approccio classico, valutato per "unità di calcolo eseguita". Abbiamo analizzato principalmente i costi relativi ai sistemi (corrente, ammortamenti), ma un analogo risultato si ha (possiamo dirlo qualitativamente e per esperienza) per le risorse sistemistiche. Nel precedente paragrafo la virtualizzazione è stata proposta come tecnologia utile per portare il costo (in vari sensi, anche di impatto sistemistico) dell'uso delle *farm* di esperimento verso quello tipico di GRID.

In questo paragrafo si tratterà di un passo ulteriore, ma sempre nella stessa direzione: l'esportazione di concetti, tecnologie e benefici di GRID in ambienti dove ancora GRID non è utilizzata. Questa attività, ed ancor più quella esposta nel precedente paragrafo, potrebbe risultare utile per un percorso di "Gridizzazione" graduale.

L'esigenza qui è centrata sull'ottimizzazione: è chiaro che la condivisione delle risorse sia la via maestra per un loro efficiente utilizzo. L'idea (non nuova, è realizzata al CERN) consiste nel dotarsi di un *pool* di macchine alle quali si rivolge un utente che necessiti di un *server* per uso interattivo (o comunque un *server* singolo con specifiche caratteristiche). Analizzando l'utilizzo della tecnologia di virtualizzazione per la realizzazione dei *server* per interattivo su richiesta siamo nello stesso alveo concettuale delle "farm on demand". Seguendo lo schema dei "driver" identificati possiamo dire:

- Obiettivi:
  - Flessibilità. È un obiettivo mutuato dalla issue della elevata dinamica, discussa di seguito
  - Efficienza. L'obiettivo è l'ottimizzazione in termini di uso efficiente delle risorse (in questo caso il minor numero medio di *server idle*). Rispetto alla situazione delle *farm* virtuali operiamo con una granularità maggiore (singolo *server*), quindi con migliori prospettive. L'efficienza è misurabile come:

$$\frac{\int_{\Delta t} (CPUworking)}{(\#CPU) \cdot \Delta t} \qquad (1)$$

e quindi dovrebbe essere presa come metro quantitativo della riuscita della attività nelle diverse sedi coinvolte (concetto valido anche per le *farm* virtualizzate).
    - Robustezza. Non è una priorità specifica in questo ambito applicativo
- Problematiche di contesto:
    - Dinamica. È un aspetto importante e risulterà di più difficile gestione (motivo per cui questo ambito applicativo viene logicamente visto successivo alla realizzazione delle "farm on demand"). L'approccio del "menu di server a la carte" risulta più complesso rispetto al caso delle *farm*. Occorrerà partire con una indagine sulle esigenze dei vari gruppi di utenti e probabilmente non sarà possibile in prima battuta prendere in carico tutte le situazioni che si presenteranno.
    - *Security*. L'aspetto della controllabilità degli utenti è analogo al caso delle *farm*: gli utenti che accedono sono singolarmente accreditati.

Anche in questo caso una promettente attività del Gruppo potrebbe essere volta ad una realizzazione pilota del sistema in oggetto in un paio di Sezioni (Parma ha già in corso una esperienza in tal senso così come Catania) che abbiano una opportuna casistica di richieste per *server* interattivi.

6 PROPOSTA DI COINVOLGIMENTO DELLA SEZIONE DI PISA

Alla luce delle esperienze fatte finora, della situazione e dei piani di Sezione relativamente al Calcolo Scientifico possiamo proporre un coinvolgimento di Pisa così articolato:

- Analisi e valutazione risorse.
    - Approfondimento degli aspetti legati all'I/O e all'impatto nella gestione sistemistica nei vari contesti. Tenendo conto della complessità della *fabric* in produzione (1G e a breve 10GE, IB InfiniBand, Myrinet, SAN con GPFS) siamo particolarmente interessati allo studio di sistemi di gestione in ottica di "fabric virtualization". Soluzioni promettenti potrebbero essere Ganeti ed Enomalism.
    - Diffusione dell'esperienza fatta con VMware sul sito GRID della Scuola Normale Superiore.
    - Aspetti legati ad una eventuale migrazione in uscita da VMware
- Virtualizzazione per *High Availability*.
    - Diffusione e documentazione dell'esperienza in corso presso la Scuola Normale Superiore.
    - Implementazione della metodologia su elementi della infrastruttura GRID di Sezione, valutazione (confronto con altre Sezioni) riguardo alla virtualizzazione di alcuni Servizi di Sezione continuando un piano di attività peraltro già avviato.

- *Farm on Demand*.
  - È in corso in Sezione una attività per la riorganizzazione delle *farm* di esperimento (recupero di efficienza). Una delle strategie, tra le altre, che stiamo prendendo in considerazione e che potrebbe essere un'interessante attività per il costituendo Gruppo riguarda:
    - individuazione di un gruppo di *farm* non utilizzate continuativamente (a Pisa abbiamo sulla carta 5 candidati). Realizzazione di un *pool* di macchine dedicate alla *delivery* di *farm* virtuali identiche a quelle di origine (mediante *cloning*) con possibilità di recupero di efficienza anche di un fattore 3-5 (inteso come rapporto tra tempo di utilizzo e tempo totale).
    - Un aspetto particolare legato a questo contesto, meritevole credo di una specifica attività, riguarda l'accesso allo *storage*. Avendo a Pisa SAN di buona qualità e robustezza, qui non si intende solo l'accesso da parte delle VM allo spazio dati, ma il mantenimento un *repository* delle VM dalla quale poter eseguire il *boot*.
- Virtualizzazione per Interattivo:
  - La creazione di un *pool* di macchine che fungano come erogatrici di risorse per gli usi interattivi è un'idea che a Pisa abbiamo da tempo, sia per specifici esperimenti che per eventuali T3. La seguente credo sia una interessante attività che ci potrebbe vedere coinvolti:
    - realizzazione e confronto di *delivery* di sistemi virtuali per interattivo secondo i due approcci:
      - a richiesta dell'utente viene lanciata sul suo *desktop* una VM come da sue specifiche;
      - sempre a richiesta si lancia la VM del tipo indicato su un *server* appartenente ad un *pool* a ciò dedicato aprendo un terminale sul *desktop* dell'utente (approccio tipo *light client*)

      I due approcci hanno diverse potenzialità e presentano potenziali difficoltà: per esempio si pensi all'accesso allo *storage* e alle periferiche locali del *desktop*. Presenta inoltre differenti problematiche di gestione sistemistica.

7 RINGRAZIAMENTI

Ringrazio:

- Silvia Arezzini (Responsabile del Servizio Calcolo e Reti di Pisa) per aver reso tale ciò che è comprensibile in queste riflessioni.
- Enrico Mazzoni, "anima" tecnica indispensabile per tentare di stare con i piedi per terra.
- Federico Calzolari (SNS) che per primo tra noi ha creduto nella virtualizzazione.

8 BIBLIOGRAFIA

(1) F. Calzolari, Proceedings of Enabling Grids for E-sciencE EGEE, Open Grid Forum OGF – 4th EGEE User Forum/OGF 25 and OGF Europe's 2nd International Event, Catania 2009.
(2) F. Calzolari, Proceedings of Computing in High Energy and Nuclear Physics CHEP 09, Prague 2009.
(3) F. Calzolari, INFN CCR Workshop, Palau 2009.