Distributed open cloud computing, storage e network con WNoDeS: Esperienza ed Evoluzione

Daniele Andreotti[1], Marco Caberletti[1], Vincenzo Ciaschini[1], Gianni Dalla Torre[1], Alessandro Italiano[2], Elisabetta Ronchieri[1], Davide Salomoni[1]

[1]INFN-CNAF, [2]INFN – Sezione di Bari

Abstract. WNoDeS è un *framework* per la virtualizzazione di risorse di calcolo che integra meccanismi di *scheduling* standard, come quelli messi a disposizione dai *batch system* utilizzati nei maggiori centri di calcolo del mondo. L'adozione di *batch system* ampiamente collaudati garantisce la scalabilità e flessibilità di WNoDeS: essa è stata verificata all'interno del centro di calcolo nazionale dell'INFN, dove è utilizzato fin dal 2009. In tale centro WNoDeS gestisce diverse migliaia di macchine virtuali create dinamicamente e messe a disposizione di comunità scientifiche internazionali. WNoDeS è inoltre integrato con soluzioni di accesso alle risorse di tipo Grid e Cloud. Questo lavoro presenta nuove funzionalità di WNoDeS, riportando alcune esperienze di utilizzo negli ultimi 4 anni e descrivendo nuove integrazioni e collaborazioni.

1. Introduzione

WNoDeS (*Worker Nodes on Demand Service*) è un framework progettato e sviluppato dall'INFN per la gestione di macchine reali e virtuali per il calcolo locale e distribuito, sia di tipo Grid che Cloud.

Le richieste architetturali che hanno portato alla realizzazione di WNoDeS includono la necessità di garantire la scalabilità, il riutilizzo di software e il bisogno di minimizzare i cambiamenti di configurazione per centri di calcolo di dimensione medio-grande. WNoDeS utilizza pertanto alcune soluzioni di virtualizzazione e *scheduling* di provata affidabilità e disponibili sul mercato: in particolare, supporta Linux KVM come virtualizzatore e IBM/Platform LSF e Torque/MAUI [1] come schedulatori. L'approccio seguito permette a WNoDeS di essere installato e configurato in un centro di calcolo per gestire risorse reali e virtuali di tipo Grid e Cloud senza richiedere agli amministratori un impegno eccessivamente oneroso. WNoDeS gestisce in particolare le risorse senza la necessità che queste siano suddivise in sottoinsiemi statici dedicati a specifiche applicazioni o interfacce, supportando trasparentemente diversi casi d'uso.

Questo articolo è strutturato come segue: la Sezione 2 descrive la struttura logica di WNoDeS; la Sezione 3 descrive una funzionalità importante di WNoDeS chiamata Mixed Mode; la Sezione IV fornisce lo stato dell'arte; la Sezione V conclude e descrive alcuni nuovi sviluppi.

2. Struttura logica

Dal punto di vista logico l'architettura di WnoDeS si può immaginare caratterizzata da un certo numero di componenti fondamentali (Figura 1), distribuite su cinque livelli.

L'*Access Layer* rappresenta il punto di accesso dell'utente al framework. Questo prevede una *Cloud Command-Line Interface* (CLI) per la sottomissione di richieste di istanziazione di macchine virtuali (VM) secondo la metodologia IaaS, e una *batch* CLI per la sottomissione di *job* di tipo *batch* o *Grid*. La Cloud CLI supporta le operazioni di creazione o cancellazione di un'istanza, di recupero di informazioni della singola istanza e di recupero di tutte le istanze associate ad un utente [5]. La *batch* CLI permette la gestione di richieste di esecuzione di *job* su macchine reali o virtuali, anche customizzate secondo le esigenze degli utenti. Per quanto riguarda

Workshop GARR Calcolo e Storage Distribuito - Selected Papers. Roma 29-30 novembre 2012
ISBN: 978-88-905077-4-8 © 2013 Associazione Consortium GARR

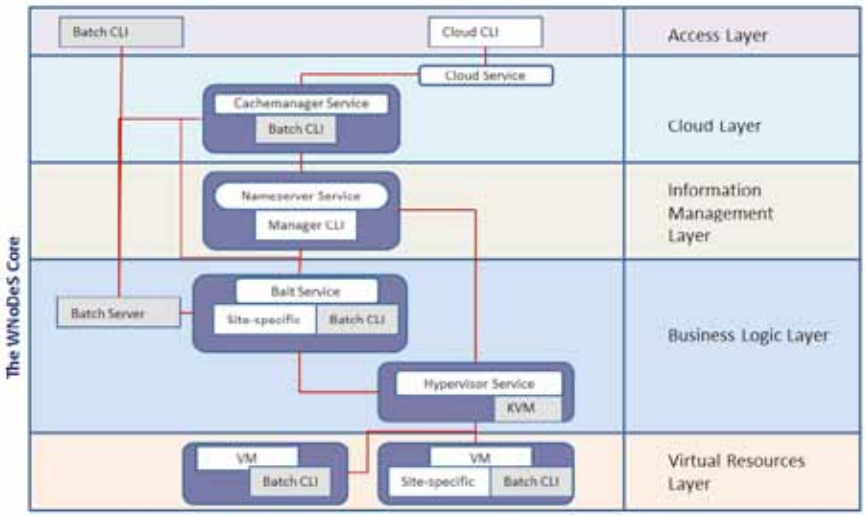

Fig. 1 - Architettura di WNoDeS

autenticazione e autorizzazione, nel caso di accessi di tipo Cloud o Grid l'utente deve appartenere a una VO e possedere un valido certificato X.509 [2]. Nel caso *batch,* l'utente deve appartenere al *pool* di utenti autorizzati dal *batch system* per sottomettere job di calcolo.

Il *Cloud Layer* [5,7] rappresenta la gestione Cloud del *framework* ed è caratterizzato da due componenti:

1. Il *Cloud Service*, che riceve le richieste di instanziazione dell'utente e le invia al *Cachemanager Service* più sotto descritto. La richiesta è marcata con un identificatore, tramite il quale il proprietario può esaminare lo stato della macchina e in particolare scoprire quando essa diventa attiva e accessibile all'utente.
2. Il *Cachemanager Service*, che gestisce la fornitura delle VM, mantenendone alcune già pronte in una cache per renderle disponibili all'utente più velocemente. La dimensione di questa cache è configurabile dall'amministratore.

L'*Information Management Layer* e il *Business Logic Layer* rappresentano la parte centrale di WNoDeS [1].

L'*Information Management Layer* è caratterizzato da due componenti, il *Nameserver Service e la Manager* CLI. Il primo può essere considerato come un catalogo che tiene traccia di tutte le VM in esecuzione su ogni *Hypervisor*, di tutte le immagini memorizzate in un opportuno repository, e delle configurazioni dei possibili *Bait* e *Hypervisor*. Il *Manager* CLI è responsabile della configurazione del *repository* delle immagini: fornisce all'amministratore una serie di opzioni per la gestione del *repository* delle immagini, delle VLAN, degli *hostname* delle immagini, e dei file di configurazione dei vari *Bait* e *Hypervisor*; fornisce inoltre una serie di opzioni per recuperare lo stato di *Bait* e *Hypervisor*.

Il *Business Logic Layer* è costituito da tre componenti:

1. L'*Hypervisor Service* è l'interfaccia al sistema di virtualizzazione responsabile dell'instanziazione delle VM (KVM). Se la richiesta utente è relativa all'esecuzione di un job (locale o Grid), tale job verrà automaticamente eseguito su una VM. Se la funzionalità *Mixed Mode* descritta nel seguito è abilitata, lo stesso *Hypervisor Service* è in grado di eseguire *batch jobs*.
2. Il *Site-specific* è il componente che permette di collegare le richieste delle risorse, internamente sempre viste come job gestiti dal batch server, al core di WnoDeS, lasciando all'amministratore la possibilità di personalizzare la richiesta dell'utente in base alle caratteristiche dell'immagine da utilizzare per l'istanziazione di una data VM. Questo componente invia inoltre la richiesta dell'utente al Bait e ne controlla lo stato.
3. Il *Bait Service* è il gestore delle risorse, responsabile di verificarne la disponibilità, di richiedere la istanziazione di VM quando necessario, e di eseguire la richiesta sulla risorsa più idonea.

Il Virtual Resources Layer rappresenta le VM istanziate sia per l'esecuzione di job tipo grid e

batch, sia per un utilizzo di tipo Cloud.

Le componenti indicate in grigio, come Batch CLI (identico nei vari *Layer* in cui è specificato come da Figura 1), *Batch Server* e Linux KVM, non sono specifici di WNoDeS.

3. Il Mixed Mode

Una delle funzionalità principali di WNoDeS permette di gestire risorse fisiche contemporaneamente sia come nodi tradizionali di un sistema batch sia come *hypervisor* per la istanziazione di VM. Tale funzionalità è detta *Mixed Mode*, è abilitabile opzionalmente e permette un'importante ottimizzazione dell'utilizzo delle risorse di un centro di calcolo, consentendo un'integrazione tra risorse reali e risorse virtuali. Come in installazioni senza *Mixed Mode*, le VM create da WNoDeS possono essere utilizzate per eseguire *job* di tipo *batch* o per fornire risorse di tipo Cloud.

Questa funzionalità permette di soddisfare alcuni requisiti che non sono comunemente gestiti da altri *framework* di virtualizzazione: ad esempio, è spesso preferibile che *job* che richiedono GPGPU o che presentano elevate richieste di I/O locale siano eseguiti senza l'*overhead* introdotto dalla virtualizzazione, e dunque su macchine fisiche. Tramite l'utilizzo di *Mixed Mode* è allo stesso tempo possibile utilizzare le medesime macchine fisiche anche per l'istanziazione di VM (purchè naturalmente risorse come memoria, disco e numero di core necessari siano sufficienti).

Il funzionamento del *Mixed Mode* è basato sul fatto che tutte le richieste di allocazione delle risorse in WNoDeS sono mediate da un *batch system*. Nel caso in cui tali richieste siano *batch job* di tipo tradizionale (sottomesse localmente o ad esempio via Grid) possiamo distinguere tra:

- *batch job* che devono essere eseguiti su un sistema fisico, senza virtualizzazione;
- *batch job* che devono essere eseguiti su VM di un certo tipo.

La distinzione tra l'uno o l'altro tipo avviene in WNoDeS attraverso un file di configurazione. Nel caso in cui le richieste di allocazione delle risorse siano di tipo Cloud, esse sono normalmente sottomesse tramite la Cloud CLI e si riferiscono alla creazione di VM. WNoDeS traduce trasparentemente queste richieste di creazione di VM in job che, come sopra, vengono passati a un batch system.

Sia che si tratti di job tradizionali che di allocazioni Cloud, dunque, per WNoDeS una richiesta di risorse è gestita da un *batch job*. Questo raggiunge un sistema fisico e in particolare il processo *Bait* di WNoDeS. Come descritto sopra, il processo *Bait* ha la responsabilità di associare da una parte le richieste di risorse associate al *job* ricevuto e dall'altra le risorse disponibili sul sistema locale. Con il *Mixed Mode* ogni sistema fisico, sul quale è in esecuzione il processo *Hypervisor* di WNoDeS, fa parte di un cluster di risorse gestito da un *batch system* e può dunque eseguire *job*. Se quindi la richiesta pervenuta al *Bait* è di esecuzione di un *job* non virtualizzato, il *Bait* manderà in esecuzione tale job direttamente sul *"bare metal"* (la macchina fisica). Se la richiesta è di istanziazione di una VM, il *Bait* richiederà all'*Hypervisor* l'istanziazione di tale VM. Se d'altra parte il *Mixed Mode* è disabilitato i sistemi fisici non faranno parte di un *batch system* e non potranno quindi eseguire job. In questo caso tutte le richieste saranno soddisfatte esclusivamente attraverso VM.

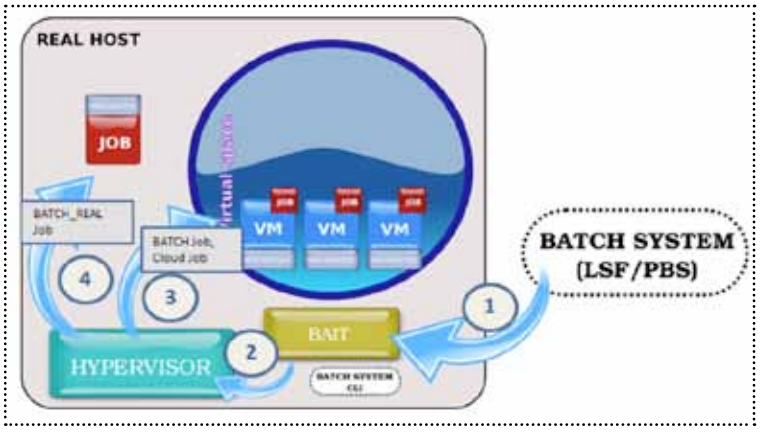

Fig 2 - Architettura di WNoDeS

Naturalmente il *Mixed Mode* richiede, per poter funzionare correttamente, che tutte le richieste di allocazione delle risorse su un certo nodo passino dal batch system e in particolare attraverso il processo *Bait*.

Il *Mixed Mode* è stato rilasciato a partire da WNoDeS 2.0.0-2 nella distribuzione EMI-2 Matterhorn [9]. Il *Mixed Mode* è inoltre integrato con il supporto fornito da WNoDeS per l'istanziazione di risorse Cloud attraverso la Cloud CLI, componente di WNoDeS rilasciata a partire a partire da WNoDeS 3.0.0-1 nella distribuzione EMI-3 Monte Bianco [9].

La figura 2 mostra il funzionamento del *Mixed Mode*, a partire da quando il *job* in coda nel sistema *batch* viene consegnato al *Bait*, che verifica lo stato delle macchine interrogando l'*Hypervisor*. In base alla tipologia di *job* e alla configurazione *site-specific*, un *job* di tipo *batch* sarà eseguito su VM o su macchina fisica. Richieste di tipo Cloud saranno invece sempre eseguite su VM.

Il *Mixed Mode* presenta vantaggi e svantaggi come descritto nella tabella seguente.

4. Stato dell'arte

In letteratura tra i *framework* di virtualizzazione e Cloud Computing di tipo *open source* sono presenti, tra gli altri, *OpenNebula* e *OpenStack*. Entrambi, come WNoDeS, creano un'infrastruttura distribuita come service (IaaS) sulla quale costruire servizi Cloud. Quello che differenzia WNoDeS da queste due tecnologie è la disponibilità della modalità *Mixed mode* e soprattutto lo sfruttamento del concetto di coda del *batch system* per la gestione delle richieste di risorse sia per job tradizionali che per allocazioni Cloud. Normalmente, infatti, le soluzioni di Cloud computing presumono una disponibilità "infinita" di risorse e implementano un sistema di *scheduling* relativamente semplice, che spesso non consente grande flessibilità nella policy di allocazione, specialmente dinamica, delle risorse (si pensi per esempio all'accodamento di richieste di allocazione, alla definizione di "*share*" di risorse, alla necessità di evitare il partizionamento delle risorse tra tenant multipli, alla scalabilità necessaria per supportare migliaia di richieste concorrenti). WNoDeS invece forni-

Vantaggi	Svantaggi
Facile installazione: rende graduale l'installazione di WNoDeS in un centro di calcolo rendendo non necessaria la preassegnazione di macchine dedicate alla virtualizzazione o al Cloud.	Il nodo reale (che contiene l'*Hypervisor*) deve essere accessibile da parte di programmi in *user-space* (job). Questa può essere una limitazione dal punto di vista della sicurezza. D'altra parte, questo è quello che accade normalmente nei casi in cui su sistemi fisici sia presente un *batch system*.
Supporto a diversi casi d'uso, in particolare ove si richieda l'allocazione flessibile di risorse reali e virtuali senza allocazione statica delle risorse.	L'utilizzo delle licenze d'uso di un *batch system* spesso aumenta proporzionalmente al numero di core visti dal *batch system* stesso. Questo può essere un problema se è disponibile solo un numero limitato di licenze (ad esempio con *batch system* di tipo commerciale). Se una macchina fisica è utilizzata per gestire *job* "reali" e per creare VM che a loro volte debbano eseguire *batch job*, il numero totale delle licenze è dell'ordine di O(2 x numero di core). Questo problema evidentemente non si pone nel caso in cui venga utilizzato un *batch system open source*.
Facile customizzazione dei *job* in base alle richieste degli utenti, che avviene tramite la modifica di un file di configurazione precompilato presente nel componente *site-specific*.	

Tab 1 - Vantaggi e svantaggi del Mixed Mode

sce, grazie alla stretta integrazione con un *batch system*, complesse politiche di allocazione e gestione risorse, comprese le componenti di autorizzazione e metering d'uso.

Un'integrazione delle architetture Cloud e Grid può anche avvenire seguendo un approccio di tipo *Grid-over-Cloud* e uno di tipo Cloud-over-Grid. I prodotti CLEVER [10] e MetaCentrum [11] implementano l'approccio *Cloud-over-Grid*. CLEVER è utilizzato per fornire un sistema IaaS realizzato a partire da una infrastruttura Grid. MetaCentrum coordina e fornisce servizi Grid nella Repubblica Ceca per conto della NGI Ceca. L'approccio *Grid-over-Cloud* è utilizzato da StratusLab [12], che fornisce servizi Grid utilizzando le risorse messe a disposizione da un'infrastruttura IaaS: la infrastruttura Grid risultante sfrutta la natura dinamica del Cloud fornendo le risorse quando necessario ed eseguendovi i servizi degli utenti oppirtunamente installati e configurati sulle risorse selezionando l'immagine della risorsa opportuna dal servizio Marketplace. Entrambi gli approcci *Grid-over-Cloud* e *Cloud-over-Grid* prevedono l'incapsulamento di una tecnologia in un'altra. D'altra parte, WNoDeS non privilegia l'interfaccia Grid oppure l'interfaccia Cloud e fornisce risorse reali o virtuali all'utente utilizzando in modo polimorfico le interfacce richieste (locali, Grid o Cloud) richieste dall'utente stesso attraverso una integrazione dinamica delle risorse.

5. Conclusioni

Da diversi anni il framework WNoDeS è utilizzato in produzione per la parte *batch* e Grid al centro di calcolo Tier-1 dell'INFN. La parte Cloud di WNoDeS, recentemente rilasciata, è attualmente messa a disposizione anche in un *testbed* fornito dalla *Italian Grid Infrastructure* (IGI) per le comunità scientifiche che ne richiedono l'accesso. Attualmente le parti specifiche di virtualizzazione Grid e Cloud di WNoDeS sono utilizzate da diversi esperimenti e collaborazioni, tra i quali si citano in particolare l'esperimento astro-particellare Auger , la infrastruttura europea WeNMR per NMR e biologia strutturale [13] e il Federated *Cloud Working Grou*p della European Grid Infrastructure (EGI) [14].

Tra le attività di prossima realizzazione sono presenti l'integrazione con il portale scientifico IGI per la parte Cloud e l'estensione della Cloud CLI per gestire l'istanziazioni di VM da parte di utenti locali per sviluppo software, test e analisi. Più a lungo termine è prevista una gestione più granulare delle risorse dei nodi attraverso meccanismi di *Control Groups* (cgroups) di Linux e l'integrazione di WNoDeS all'interno di OpenStack, utile soprattutto per permettere allo stesso OpenStack di supportare *workload* di calcolo scientifico in modo scalabile. All'interno di quest'ultima attività sono preliminarmente previste, in particolare, l'integrazione dello *scheduling* gestito da WNoDeS con lo scheduling di OpenStack, l'estensione della *dashboard* di Openstack per l'amministrazione ed il monitoraggio dei servizi di WNoDeS e l'integrazione del servizio prototipale di *Dynamic Virtual Networks* di WNoDeS [8] all'interno della gestione rete di OpenStack.

Riferimenti Bibliografici

[1] Davide Salomoni, Alessandro Italiano, Elisabetta Ronchieri, "WNoDeS, a Tool for Integrated Grid and Cloud Access and Computing Farm Virtualization," 2011 Journal of Physics: Conference Series Volume 331 Part 5: Computing Fabrics and Networking Technologies.

[2] Vincenzo Ciaschini, Davide Salomoni, "An Authentication Gateway for Integrated Grid and Cloud Access," 2011 Journal of Physics: Conference Series Volume 331 Part 6: Grid and Cloud Middleware.

[3] Claudio Grandi, Alessandro Italiano, Davide Salomoni, Anna Karen Calabrese Melcarne, "Virtual Pools for Interactive Analysis and Software Development through an Integrated Cloud Environment," 2011 Journal of Physics: Conference Series Volume 331 Part 7: Distributed Processing and Analysis.

[4] Davide Salomoni, Anna Karen Calabrese Melcarne, Andrea Chierici, Luca Cestari, Gui-

do Potena, Peter Solagna "Performance improvements in a large scale virtualization system," PoS(ISGC 2011 & OGF 31)049.

[5] Davide Salomoni, Daniele Andreotti, Luca Cestari, Guido Potena, Peter Solagna, "A Web-based Portal to Access and Manage WNoDeS Virtualized Cloud Resources," PoS(ISGC 2011 & OGF 31)054.

[6] Davide Salomoni, Elisabetta Ronchieri, "WORKER NODES ON DEMANDS SERVICE – Requirements for Virtualized Services, " http://web2.infn.it/wnodes/index.php/documentation/files-download/28_d67514b33dae20f979d866990b583b74

[7] Elisabetta Ronchieri, Giacinto Donvito, Paolo Veronesi, Davide Salomoni, Alessandro Italiano, Gianni Dalla Torre, Daniele Andreotti, Alessandro Paolini, "Resource Provisioning through Cloud and Grid Interfaces by means of the Standard CREAM CE and the WNoDeS Cloud Solution," PoS(EGICF12-EMITC2)124.

[8] Marco Caberletti, Davide Salomoni, "A Dynamic Virtual Networks Solution for Cloud Computing," Proceedings of the 2nd International Workshop on Netware-aware Data Management, November 11, 2012, Salt Lake City, Utah, USA.

[9] Cristina Aiftimiei, Andrea Ceccanti, Danilo Dongiovanni, Andrea Di Meglio, Francesco Giacomini, "Improving the quality of EMI Releases by leveraging the EMI Testing Infrastructure," 2012 Journal of Physics: Conference Series Volume 396 Part 5.

[10] Francesco Tusa, Maurizio Paone, Massimo Villari, Antonio Puliafito, "CLEVER: a Cloud Cross-Computing Platform leveraging GRID resources," UCC pp. 390 – 396, IEEE Computer Society (2011).

[11] Ruda Miroslav, Šustr Zdenek, Sitera Jirí, Antoš David, Hejtmánek Lukáš, Holub Petr, "Virtual Clusters as a New Service of MetaCentrum, the Czech NGI," In Cracow Grid Workshop '09. Krakow : Academic Computer Centre CYFRONET AGH, 2010. ISBN 978-83-61433-01-9, pp. 64-71. 12.10.2009, Krakow.

[12] Charles Loomis, Mohammed Airaj, Marc-Elian Bégin, Evangelos Floros, Stuart Kenny, David O'Callaghan, "StratusLab Cloud Distribution," In Dana Petcu and José Luis Vázquez-Poletti Eds., European Research Activities in Cloud Computing, pp. 260–282, Cambridge Scholars Publishing (2012).

[13] Tsjerk A. Wassenaar, Marc van Dijk, Nuno Loureiro-Ferreira, Gijs van der Schot, Sjoerd J. de Vries, Christophe Schmitz, Johan van der Zwan, Rolf Boelens, Andrea Giachetti, Lucio Ferella, Antonio Rosato, Ivano Bertini, Torsten Herrmann, Hendrik R. A. Jonker, Anurag Bagaria, Victor Jaravine, Peter Güntert, Harald Schwalbe, Wim F. Vranken, Jurgen F. Doreleijers, Gert Vriend, Geerten W. Vuister, Daniel Franke, Alexey Kikhney, Dmitri I. Svergun, Rasmus H. Fogh, John M. C. Ionides, Ernest D. Laue, Chris A. E. M. Spronk, Simonas Jurksa, Marco Verlato, Simone Badoer, Stefano Dal Pra, Mirco Mazzucato, Eric Frizziero, Alexandre M. J. J. Bonvin, "WeNMR: Structural Biology on the Grid," J. Grid. Computing. V. 10, pp. 743-767.

[14] Matteo Turilli, Michel Drescher, Steven Newhouse, David Wallom "Federeting clouds to aid researchers", ISGTW – International Science grid this week, 17 October 2012.

Elisabetta Ronchieri
elisabetta.ronchieri@cnaf.infn.it

Lavora presso il CNAF dell'INFN dal 2001, dopo aver trascorso un paio di anni nell'industria sviluppando software di carte nautiche e simulazioni di sistemi dinamici. Ha partecipato a numerosi progetti Europei riguardanti software engineering e calcolo distribuito in ambito Grid e Cloud. Attualmente gestisce il progetto WNoDeS, fa parte del Fedcloud Working Group del progetto Europeo EGI Inspire e collabora allo sviluppo di modelli per la valutazione della qualità del software.