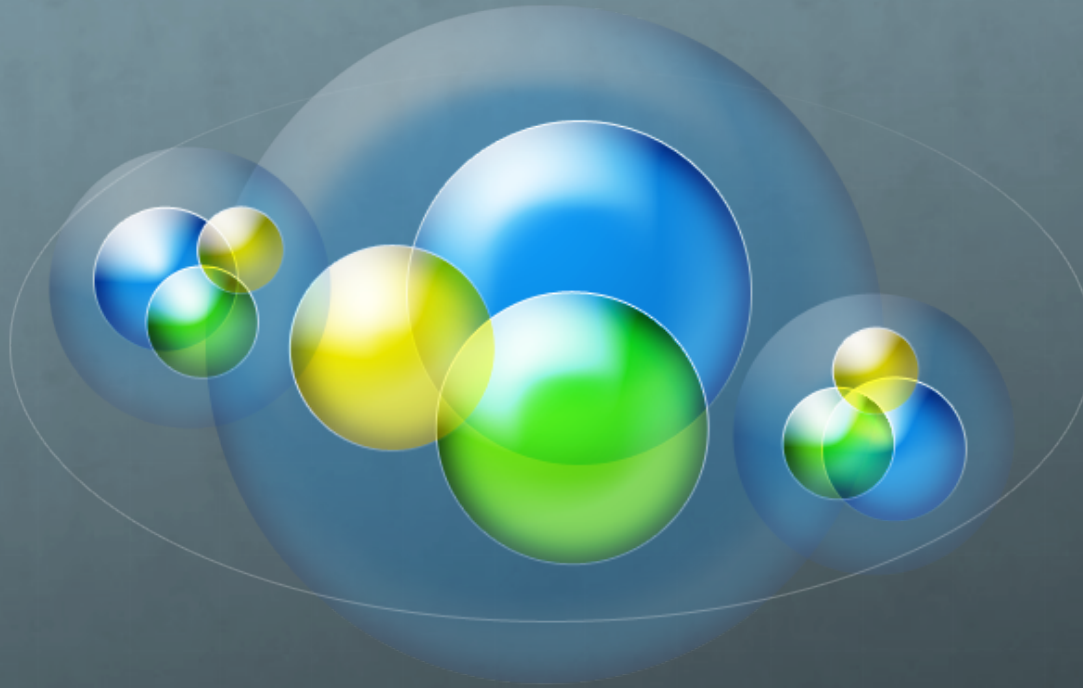




A short introduction to grid and cloud computing

Maarten Litmaath (CERN)
Alessandro De Salvo (INFN Roma/CERN)
Alessandro.DeSalvo@roma1.infn.it

Outline

- **What is a grid?**
- **Are grids obsoleted by clouds?**
- **What are grids about?**
- **How does a grid work?**
- **Where can we use grids?**
- **What is out there?**
- **A concrete example of running grids in a HEP experiment: ATLAS**
- **Conclusions**

What is a grid?

- Relation to WWW?
 - Uniform easy access to shared information
- Relation to distributed computing?
 - Local clusters
 - WAN (super)clusters
 - **Condor**
- Relation to distributed file systems?
 - NFS, AFS, GPFS, Lustre, PanFS...



- A grid gives selected user communities uniform access to distributed resources with independent administrations

– Computing, data storage, devices, ...

Why is it called grid?

- Analogy to power grid
 - You do **not need to know** where your electricity comes from
 - Just plug in your devices
- You should **not need to know** where your computing is done
 - Just plug into the grid for your computing needs
- You should **not need to know** where your data is stored
 - Just plug into the grid for your storage needs



Ian Foster's checklist

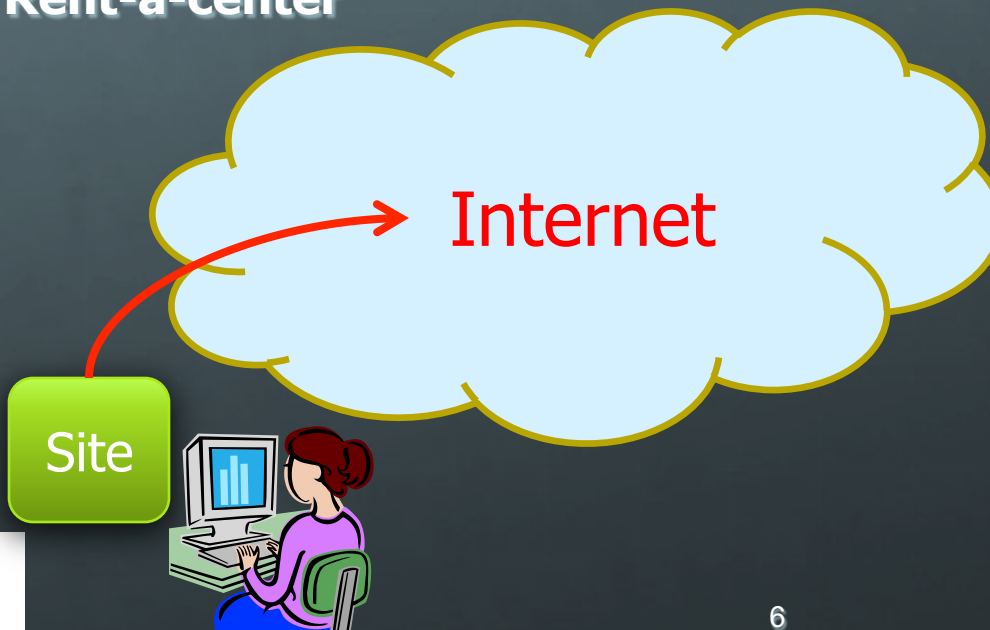
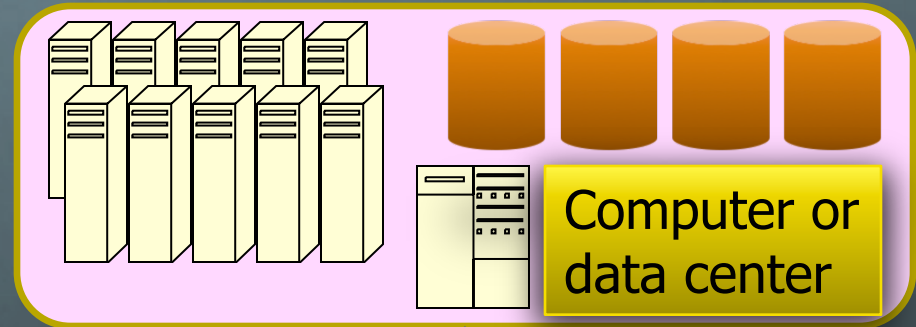
<http://www-fp.mcs.anl.gov/~foster/Articles/WhatIsTheGrid.pdf>

Globus

- A grid coordinates resources that are not subject to centralized control,
- using standard, open, general-purpose protocols and interfaces,
- to deliver non-trivial qualities of service.
 - Response time
 - Throughput
 - Capacity
 - Availability
 - Security
 - Co-allocation of multiple resource types for complex work
- Utility of the combined system significantly greater than the sum of its parts

What is cloud computing?

- Transparent use of generic computing resources off-site
 - Dynamically provisioned
 - Metered
- Neutral to applications
 - Rent-a-center



- Amazon EC2
- Amazon S3
- Sun
- ...

Cloud computing

- La definizione classica di riferimento è quella del National Institute of Standards and Technology (NIST) USA (<http://goo.gl/eBGBk>)
- In sintesi il Cloud computing si occupa di:

Fornitura di tecnologia di informazione
e comunicazione (ICT) come servizio

Caratteristiche del Cloud

- **Self-service, on-demand**
 - Il cliente chiede autonomamente ciò che gli serve, quando gli serve (e sperabilmente lo ottiene).
- **Accesso attraverso la rete**
 - Assume che una rete (Internet o intranet) sia disponibile, normalmente a banda larga.
- **Pool di risorse**
 - L'utente non si preoccupa di conoscere i dettagli delle risorse, che sono gestiti dai Cloud resource provider.
- **Elasticità**
 - Il servizio Cloud può scalare rapidamente come dimensioni a seconda delle necessità del cliente.
- **Pagamento a consumo**
 - Il cliente paga solo per ciò che usa.

Il focus sul “service”

- Abbiamo visto che nella definizione di Cloud computing (“Fornitura di tecnologia di informazione e comunicazione come servizio”) il ***servizio*** nei confronti del cliente è parte essenziale.
- Il Cloud computing si può modellare infatti intorno a *servizi* legati principalmente a
 - Infrastruttura (**IaaS** → Infrastructure as a Service)
 - Piattaforma (**PaaS** → Platform as a Service)
 - Software (**SaaS** → Software as a Service)

Deployment: i “tipi di Cloud”

- **Cloud privata:**
 - L’infrastruttura viene fornita per un *uso* esclusivo da parte di una singola organizzazione. La gestione, l’operazione, la proprietà, la dislocazione della Cloud privata tuttavia può essere anche indipendente dall’organizzazione che la usa.
- **Cloud di comunità (Community Cloud):**
 - L’infrastruttura è disponibile ad una comunità di organizzazioni che hanno uno scopo comune (ad esempio missione, requisiti di sicurezza, conformità a regole comuni, etc.)
- **Cloud pubblica:**
 - L’infrastruttura è disponibile in generale al pubblico. La gestione può essere pubblica o privata. La dislocazione è presso il fornitore di servizi.
- **Cloud ibrida:**
 - L’infrastruttura è una combinazione di due o più infrastrutture Cloud (private, di comunità o pubbliche) che sono collegate in modo da garantire forme di portabilità ad esempio di dati o applicazioni.

Isolamento

- I modelli di **isolamento** nel Cloud (spesso ignorati) sono importanti e si dividono in:
 - Infrastrutture dedicate
 - Infrastrutture “multi-tenant” (con diversi [tipi di] clienti)
- Il tipo di isolamento è importante per molti aspetti, come:
 - Segmentazione delle risorse
 - Protezione dei dati
 - Sicurezza delle applicazioni
 - Auditing
 - Disaster recovery

Virtualizzazione

- Il Cloud computing può anche essere fornito *senza* l'utilizzo di tecnologie di virtualizzazione.
 - Spesso, tuttavia, l'utilizzo di tecnologie di virtualizzazione consente di ridurre i costi operativi e in conto capitale.
 - Essere in grado di fornire molto rapidamente delle macchine virtuali non è comunque efficiente, se ad esempio servono diversi mesi per acquisire ed installare gli host fisici che ospiteranno le macchine virtuali.
 - Inoltre, il tempo impiegato per la fornitura o gestione dello strato di virtualizzazione è recuperato dai risparmi associati al non dover dedicare dei server fisici?
 - Importanza di avere tool di installazione, monitoring e accounting il più possibile automatizzati.
- Ma che cosa si intende con *virtualizzazione*?

Vantaggi della virtualizzazione

(1/2)

- Consolidamento di server.
 - Molte VM su una macchina fisica.
 - Riduce i costi di acquisizione dell'hardware e può semplificare operazioni come monitoring e amministrazione.
- Isolamento (*sandboxing*).
 - Isolamento delle applicazioni.
 - Sviluppo, test e debug di codice.
 - Creazione di ambienti dedicati per applicazioni legacy.
- Creazione di VM *on-demand*.

Vantaggi della virtualizzazione

(2/2)

- Disaccoppiamento dell'hardware dal software.
 - Suspend/Resume delle VM.
 - Migrazione di VM tra server fisici (con vari gradi di sofisticazione).
- Test di nuove versioni di sistema operativo, applicazioni.
 - O di vecchie versioni: data preservation.
- Emulazione di hardware non presente sulla macchina fisica.
- Esecuzione di applicazioni non compatibili con il sistema operativo della macchina fisica.

Svantaggi della virtualizzazione

- Sicurezza.
 - Si mettono sullo stesso hardware diversi sistemi operativi, gestiti da un software – aumento della probabilità di bug e aumento del numero dei cosiddetti *attack vectors*:
 - VM-to-VM network attacks.
 - VM-to-HV (KVM o XEN). KVM è un modulo del kernel Linux. Xen è direttamente connesso all'hardware. → tutto può diventare compromesso.
 - VM-to-QEMU. QEMU è un software grande e complesso. Se viene compromesso, l'attacco può arrivare al sistema operativo.
- Performance.
 - Overhead per la macchina fisica.
 - Minori prestazioni per la VM, soprattutto nell'I/O.

Riassumendo: virtualizzazione vs. Cloud computing

- **Installazione/reinstallazione** di server o di applicazioni su VM di per sé ***non è Cloud computing***.
- Verificare con le **5 caratteristiche del Cloud** mostrate precedentemente:
 - Self-service, on-demand → **NO** (tipicamente è un dipartimento IT che fornisce le VM)
 - Accesso attraverso la rete → **NO** (deployment limitato a “internal customers”)
 - Pool di risorse → **SÌ**
 - Elasticità → **NO** (tipicamente è un dipartimento IT che deve installare sistema operativo + software, e non necessariamente in modo scalabile)
 - Pagamento a consumo → **NO** (spesso il billing non viene fatto a consumo ma in modo tradizionale)

Integrazione tra virtualizzazione e Cloud computing

- Nonostante le differenze, ci sono chiaramente **punti di sovrapposizione** tra framework di virtualizzazione e framework di Cloud computing.
- Nel mondo open source, un caso di integrazione è dato dalla crescente **convergenza tra oVirt (<http://www.ovirt.org/>) e OpenStack**.
 - “With [oVirt 3.3], the oVirt community begins integrating oVirt with other important open source projects such as OpenStack and Gluster.” (I. Hein, RedHat)
 - Integrazione delle immagini (Glance) e del networking (Neutron) forniti da OpenStack.
 - Integrazione di cloud-init, il tool standard *de facto* utilizzato per la contestualizzazione di nodi Cloud.

Tipi di virtualizzazione

- Quando i sistemi virtuali usano un insieme di istruzioni di linguaggio macchina (instruction set) *differente* da quello disponibile sul sistema ospitante, si parla di **emulazione**.
 - Esempio: emulare ARM o PowerPC su piattaforma Intel.
- Quando i sistemi virtuali usano *lo stesso* instruction set del sistema ospitante, si parla propriamente di **virtualizzazione**.
 - A sua volta distinta tra *virtualizzazione completa* e *para-virtualizzazione*.

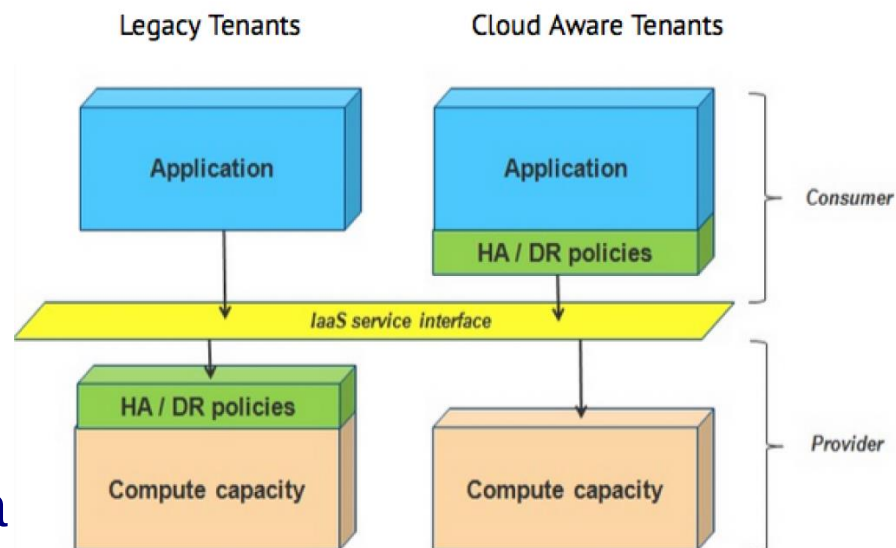
La mia applicazione è “cloud-friendly”?

- **Applicazioni “cloud-aware”:**

- Distribuite
- Stateless
- Fail-over in the app
- Scaling in the app

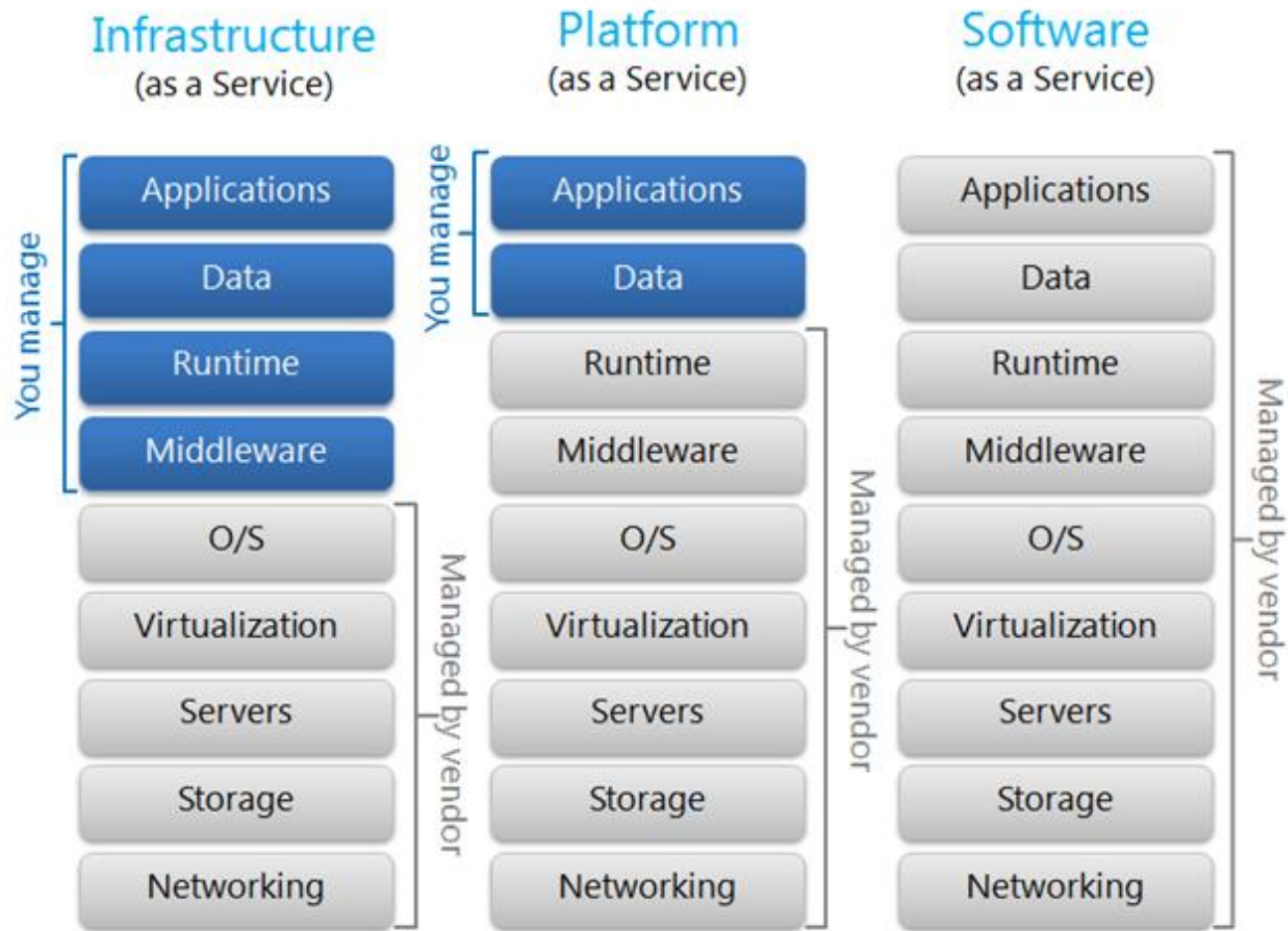
- **Applicazioni “legacy”:**

- Client-server
- Monolitiche, senza scalabilità orizzontale
- Fail-over nell’infrastruttura
- Scaling nell’infrastruttura



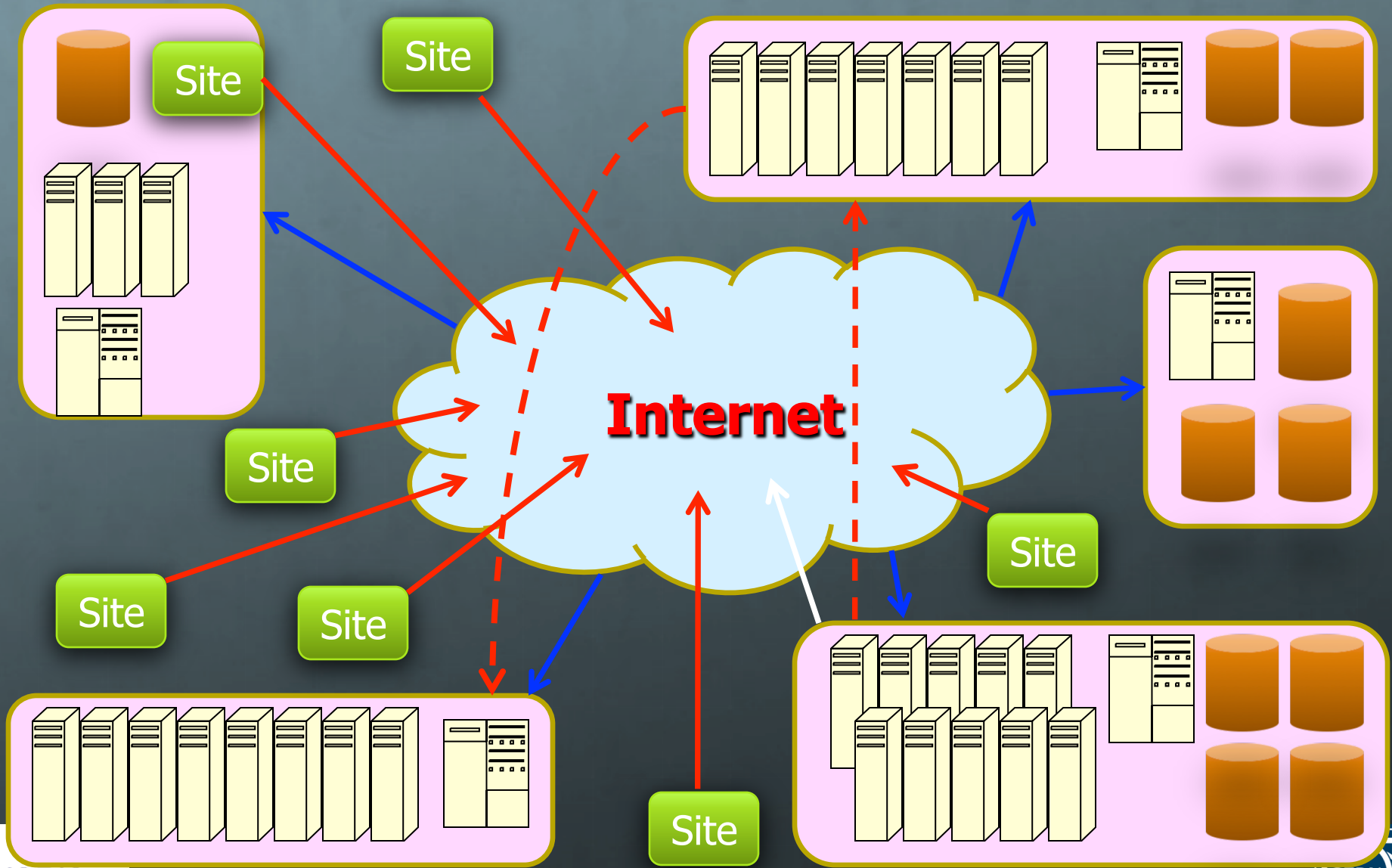
Fonte: VMware

Chi fa cosa?



Fonte: <http://goo.gl/1jmkR>

What is grid computing?



What is grid computing about?

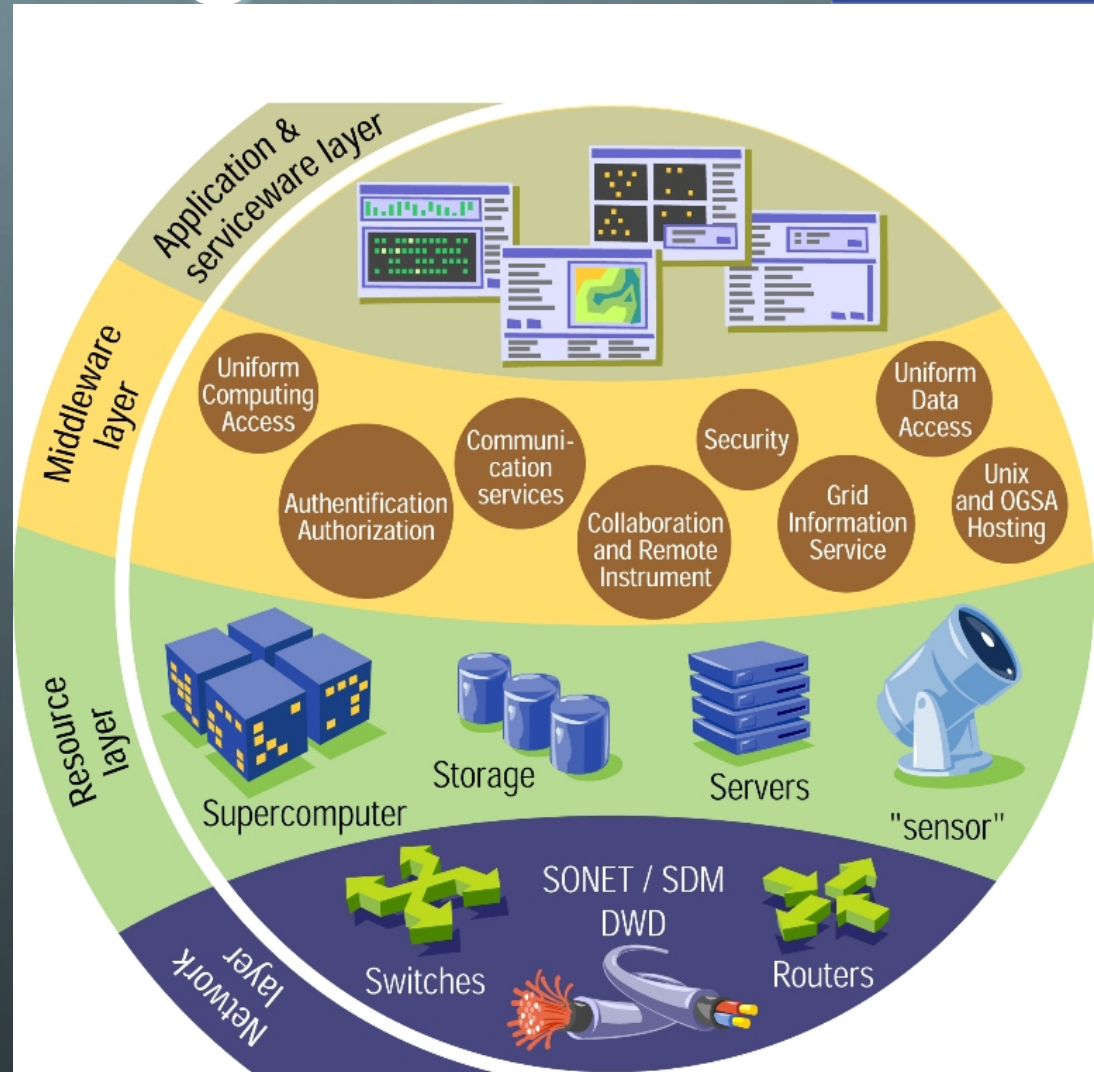
-  A grid facilitates collaboration between members of a supported distributed community
 -  They can form a **Virtual Organization** within that grid

-  A grid allows distributed resources to be shared uniformly and securely for common goals
 -  Computing
 -  Data storage

-  A grid can support multiple Virtual Organizations in parallel
 -  Sites, computer and data centers make selections according to the projects in which they participate
 -  The quality of service may differ per VO

How does a grid work?

- **Middleware** makes multiple computer and data centers look like a single system to the user
- **Security**
- **Information system**
- **Data management**
- **Job management**
- **Monitoring**
- **Accounting**
- **Not easy!**



Conclusions

- Grids facilitate collaboration between members of supported distributed communities
- Grids allow distributed resources to be shared uniformly and securely for common goals
- Grids may have complex infrastructures
- Grid and cloud computing are an unprecedented step forward with respect to traditional computing approaches when a big amount of data must be handled, and when the reliability and availability are the main issues
- Grids are useful for many scientific disciplines and projects
- Cloud computing is more targeted to commercial activities, while keeping the underlying structure very similar to grids and mostly using the same tools to give uniform access to resources