**PON 254 – “Beni Culturali”, Università di Salerno**

**Architettura della Cloud**

**Versione 0.1 (18/7/2011)**

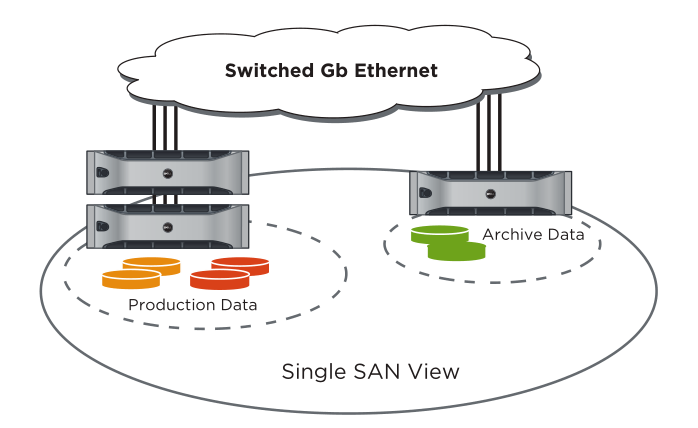
# Linee guida del progetto

#### Scalabilità verticale e orizzontale

La scalabilità è un requisito fondamentale nelle infrastrutture che hanno il compito di gestire i contenuti non strutturati per molteplici applicazioni e reparti aziendali. Con un infrastruttura scalabile è possibile gestire grandi volumi di contenuti e soddisfare requisiti di prestazioni superiori. I requisiti di scalabilità si applicano quindi a tutti i componenti dell'infrastruttura, inclusi il repository, i servizi per i contenuti e le applicazioni che li utilizzano.

In particolare per sistemi di storage massivo, la scalabilità può essere intesa in due direzioni: si parla di scalabilità verticale, quando si esamina la possibilità di ampliare la capacità di memorizzazione (o di elaborazione o di memoria RAM) del sistema, aggiungendo componenti al sistema; si intende, invece, per scalabilità orizzontale la caratteristica di poter offrire prestazioni elevate per quanto riguarda la distribuzione di carico, intendendo, quindi, ad esempio per la memorizzazione, la distribuzione di dati in maniera bilanciata su tutti i sottosistemi di storage in modo da poter utilizzare al massimo la velocità di accesso al disco e la velocità di comunicazione con la infrastruttura di calcolo.

#### Strutturazione di memoria a livelli

Le Storage Area Network (SAN) hanno come obiettivo il trasferimento di dati tra sistemi di computer ed elementi di storage e tra elementi di storage. Una rete SAN consiste in un’infrastruttura di comunicazione, che fornisce connessioni fisiche e in un livello di gestione, che organizza connessioni, elementi di storage e sistemi di computer in modo da garantire un trasferimento di dati sicuro e robusto. Se tutti i carichi di lavoro nell'ambiente fossero di uguale importanza e tipologia, una strutturazione omogenea di storage si può efficientemente adattare, ma spesso (ed è il caso delle applicazioni considerate) i carichi di lavoro  
possono avere diversi livelli di importanza e di caratterizzazione. La risposta progettuale è quella di prevedere diversi livelli (tier) di memorizzazione di massa, con obiettivi diversi, in particolare, lo spazio di “production” (dove la necessità è assicurare il massimo throughput verso il sistema di calcolo con la parallelizzazione massiva e la scalabilità orizzontale) e lo spazio di “archive” dove l’obiettivo, invece, diventa massimizzare lo spazio a disposizione a parità di costo. In un tale sistema, gli amministratori possono configurare diversi datastore separati in pool, all’interno di una singola SAN per ottenere il massimo di flessibilità e di efficacia dell’investimento, mantenendo la possibilità di gestire picchi di carico secondo le esigenze.

#### Virtualizzazione dei servizi offerti e delle macchine di gestione e controllo

Con la virtualizzazione, una macchina, di solito un server, viene multiplexata da un piccolo kernel con priviligi, tipicamente chiamata hypervisor o monitor, che fornisce la illusione di più macchine reali fisiche. Questo meccanismo è stato originariamente utilizzato per permettere a più sistemi operativi di coesistere in modo sicuro su una macchina fisica. Ma successivamente la virtualizzazione nelle imprese ha avuto come obiettivo quello di ridurre i costi mediante l'esecuzione dei carichi di lavoro come macchine virtuali (VM), su un minor numero di server in grado di migliorarne l’uso, in modo da poter utilizzare (al crescere del carico) le stesse risorse che non sarebbero state sufficienti senza l’uso di virtualizzazione.

La virtualizzazione è utile nel migliorare la qualità della gestione delle infrastrutture IT. Infatti, i server virtuali possono essere migrati “a caldo” senza tempi morti, senza la necessità di un coordinamento con gli utenti per gli aggiornamenti hardware.

La virtualizzazione avviene tramite un cosiddetto processo di consolidamento che mira a ridurre il numero di server sottoutilizzati dalla coabitazione di più macchine virtuali su un unico server fisico: questo permette di liberare macchine fisiche che possono essere a disposizione per ottimizzazioni sul carico di lavoro (load balancing) o per politiche di energy-saving.

Il sistema proposto si appoggia alla virtualizzazione per due aree specifiche:

1. Macchine di management e gestione, con vincoli di real-time non stringenti, possono essere virtualizzate in modo da utilizzare al meglio le risorse disponibili, in modo da lasciare al calcolo ad alte prestazioni la massima parte possibile delle risorse.
2. Macchine per servizi in hosting, che vengono eseguite per commesse esterne, in modo da facilitare l’utilizzo da parte di utenti esterni. Infatti, la macchina virtuale “tipo” può essere fornita dal centro al potenziale cliente che può sviluppare, in locale sulla macchina virtuale, il tipo di applicazione per le analisi che intende compiere, su un data set campione. Al termine del testing, il client può sottomettere la macchina virtuale al sistema, che eseguirà la applicazione sui dati indicati, memorizzati nel sistema.

# Architettura di massima

## Componenti del sistema

L’infrastruttura Hardware comprende

* Nodi di calcolo ad alte prestazioni basati su sistemi Blade
* Nodi di calcolo ibridi basati su sistemi Modulari ad Alta Densità interconnessi a GPU
* Rete di interconnessione tra i nodi Blade a bassa latenza basata su tecnologia InfiniBand DDR/QDR
* Rete Locale di Interconnessione basata su tecnologia full-10Gb Ethernet, in topologia dual-core con interconnessione diretta tra i nodi ed il core (direct-attach)
* Rete locale di management basata su tecnologia 1Gb Ethernet in topologia core-edge, con singolo core interconnesso a switch “edge” posizionati in ogni rack
* Storage farm basata su protocollo iSCSi dotati di tecnologia peer-storage a scalabilità orizzontale per l’accesso block-level mediante protocollo iSCSI e su sistemi gateway per la connettività di tipo File System su rete 10GBEth

### Nodi di Calcolo Tradizionale basati su Blade

L’organizzazione tipica dei sistemi di calcolo intensivo si basa sui sistemi “Blade”, che consistono di schede (“Blade”, lame) multiprocessore, inserite in uno chassis, a sua volta localizzato in un rack.

I nodi di calcolo sono fisicamente basati su nodi dual-socket (quad-socket in prospettiva futura), con architettura INTEL Nehalem. Ogni nodo è dotato di 2GB di RAM per core fisico con un minimo di 24GB di RAM, e “dialoga” verso l’esterno (DC LAN e Storage) mediante 4 interfacce di rete 10GbEth “converged” direttamente connesse al centro stella in direct-attach tramite moduli “passtrough”

La connessione a bassa latenza tra i nodi è realizzata invece mediante 2 connessioni IB DDR a 20Gbps attestate sulle due fabric IB interne allo chassis, a loro volta attestate in uplink verso lo switch IB.

Si noti che sui nodi è possibile eseguire i più diffusi SW di virtualizzazione: ciò permette di impiegare l’infrastruttura di calcolo anche per scopi differenti dal mero calcolo ad alte prestazioni.

### Nodi di Calcolo Ibridi basati su sistemi Modulari ad Alta Densità interconnessi a GPU

Fisicamente basati su Server modulari ad alta densità dotati cadauno di 4 nodi dual-socket, ed architettura INTEL Nehalem. Ogni nodo è dotato di 2GB di RAM per core fisico con un minimo di 24GB di RAM, e “dialoga” verso l’esterno (DC LAN e Storage) mediante 2 interfacce di rete 10GbEth direttamente connesse agli switch Ethernet di centro stella. Ogni nodo è interconnesso mediante un sistema di estensione del bus PCIe ad uno chassis C4100 contenente 16 moduli GPU basati su tecnologie Nvidia Tesla.

Per la massima flessibilità è possibile svincolare le GPU dal nodo di calcolo (rendendo i nodi di calcolo autonomi) ed anche riorganizzare la distribuzione dei nodi GPU rispetto ai nodi di calcolo (assegnando ad esempio fino a 4 sistemi GPU ad un singolo nodo o nessun sistema GPU ad altri nodi.

### Rete di Interconnessione a Bassa Latenza per i nodi di calcolo

Basata su protocollo InfiniBand, garantisce la comunicazione a bassa latenza tra i nodi attraverso una mesh IB con connessioni DDR/QDR. Fisicamente costituita dai moduli IB switch interni ai sistemi Blade a loro volta attestati in uplink su Director di classe enterprise.

### Rete Locale di Interconnessione (10GbEth)

Garantisce la comunicazione e la movimentazione dei dati tra i nodi e le altre componenti del sistema di calcolo e degli altri sistemi esistenti (“resto del mondo”) attraverso 2 o 4 interconnessioni 10gbEth per sistema afferenti direttamente ad un centro stella ridondato (doppio core).

E’ fisicamente basata su tecnologia “port blades” da 10Gbps, per una scalabilità massima di 640 porte 10GbEth per sistema in “shared bandwidth” (oversubscription ratio 1:4).

### Rete Locale di Management

Implementa tramite una rete separata la possibilità di effettuare il management di tutte le componenti dell’infrastruttura. La topologia di implementazione è di tipo core-edge, con i core condivisi con quelli della rete di Interconnessione di cui al punto precedente (verranno utilizzate alcune porte a 10GbEth sui core) e gli edge costituiti da switch a 48 porte 1GbEth “in testa di rack” interconnessi con 2 link a 10gbEth verso i 2 core e con link a 1Gbps verso i sistemi presenti nel rack.

### Storage Farm

Implementa l’area di memorizzazione dei dati mediante protocollo di accesso a blocchi di tipo iSCSI attraverso la rete Ethernet /IP.

E’ fisicamente composto da un pool di sottosistemi a disco per memorizzazione di “secondo livello”, cadauno dotato di 48 dischi da 2TB a 7.2krpm (per un totale di 96 TB RAW) e 4 porte 10GbEth (2 per controller) e di un sistema di memorizzazione di “primo livello”, cadauno dotato di 16 dischi da 600GB a 15krpm (per un totale di 9,6 TB RAW) e 4 porte 10GbEth (2 per controller).

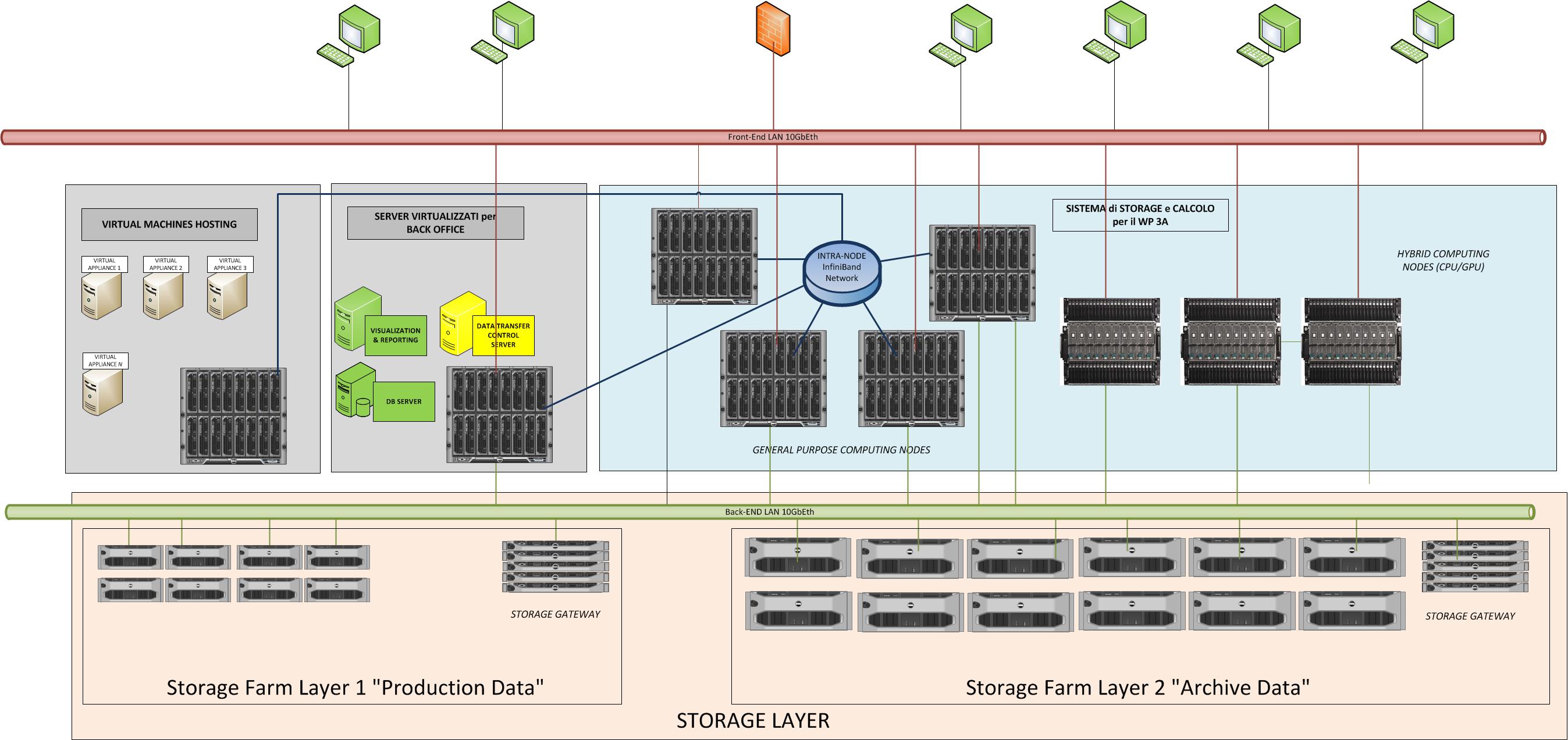
Grazie alla tecnologia “peer storage” di cui è dotato il prodotto, vengono eliminati i potenziali colli di bottiglia del sistema e si realizza la desiderata scalabilità orizzontale del layer di memorizzazione: in pratica ogni volume (LUN) è diviso in blocchi, che sono equamente ed automaticamente distribuiti su tutti i nodi disponibili. Quindi, nella media, tutti gli accessi sono equamente distribuiti su tutti i nodi e tutti i nodi EQL gestiscono lo stesso carico medio attraverso tutte le interfacce attive. Pertanto anche un solo nodo di calcolo può parallelizzare il carico su più macchine in back/end.

La connettività di tipo File System è invece implementata attraverso NAS Gateway, che permettono di ottenere un sistema Gateway a scalabilità orizzontale.

# Configurazione del nodo singolo

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Sottosistema*** | ***Nome*** | ***Descrizione*** | ***Costo*** | ***IVA*** | ***Costo unitario*** | ***Unità*** | ***Totale*** | ***CPU(#)*** | ***Core(#)*** | ***GPU(#)*** | ***RAM(GB)*** | ***Dischi (Gb)*** |
| ***Sistema in toto*** |  |  |  |  |  |  | ***€ 2.227.064*** | ***68*** | ***272*** | ***16*** | ***2560*** | ***535,2*** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| Nodi di calcolo tradizionale basati su Blade |  |  |  |  |  |  | **€ 301.829** |  |  |  |  |  |
|  | Chassis per 16 blade | Controller, KVM controller, alimentazione ridondante, moduli di rete 10Gbs ridondati, Switch Infiniband (1+2) | € 16.250 | € 3.250 | € 19.500 | 2 | **€ 39.000** |  |  |  |  |  |
|  | Blades | Dual network board 10Gb, bi processori classe Xeon (o superiori) quadcore, con hyperthreading , 16Mb cache, 64Gb RAM, Scheda Infiniband, Software di gestione | € 6.845 | € 1.369 | € 8.213 | 32 | **€ 262.829** | 64 | 256 |  | 2048 |  |
| Nodi di calcolo ibridi basati su sistemi modulari ad alta densita interconnessi a GPU |  |  |  |  |  |  | **€ 103.935** |  |  |  |  |  |
|  | Server ad alta densità | Alimentazione ridondata, 2 processori quad-core classe Xeon (o superiore) con 16 Mb cache, hyperthreading, 256Gb RAM, 4 GPU nVIDIA, 12 dischi da 300Gb | € 39.000 | € 7.800 | € 46.800 | 2 | **€ 93.600** | 4 | 16 |  | 512 | 7,2 |
|  |  | Espansione PCIExpress per 16 schede GPU nVidia, con 16 schede | € 8.613 | € 1.723 | € 10.335 | 1 | **€ 10.335** |  |  | 16 |  |  |
| Switch di centro stella |  |  |  |  |  |  | **€ 702.900** |  |  |  |  |  |
|  | Chassis | Alimentatori ridondanti, 16 slot per port blade | € 68.250 | € 13.650 | € 81.900 | 1 | **€ 81.900** |  |  |  |  |  |
|  | Port blade | 40 porte 10Gbs | € 32.500 | € 6.500 | € 39.000 | 15 | **€ 585.000** |  |  |  |  |  |
|  | Switch | 162 porte Infiniband a 40Gb/s per port | 30000 | € 6.000 | € 36.000 | 1 | **€ 36.000** |  |  |  |  |  |
| Per rack: |  |  |  |  |  |  | **€ 249.600** |  |  |  |  |  |
|  | Switch in "testa di rack" | 48 porte 1 Gb e 2 porte 10Gbs | 41600 | € 8.320 | € 49.920 | 5 | **€ 249.600** |  |  |  |  |  |
| Storage Farm Layer 1 "Production Data" |  |  |  |  |  |  | **€ 309.600** |  |  |  |  |  |
|  | Unità di storage | 16 dischi x 600GB SAS-II, velocità da 15K, doppio controller | € 41.600 | € 8.320 | € 49.920 | 5 | **€ 249.600** |  |  |  |  | 48 |
|  | NAS Gateway e Management | FS interface, peer decoupling, RAID, backup automatico, recovery, management | € 10.000 | € 2.000 | € 12.000 | 5 | **€ 60.000** |  |  |  |  |  |
| Storage Farm Layer 2 "Archive Data" |  |  |  |  |  |  | **€ 559.200** |  |  |  |  |  |
|  | Unità di storage | 48 dischi x 2TB SATA da 7.2K, doppio controller | € 83.200 | € 16.640 | € 99.840 | 5 | **€ 499.200** |  |  |  |  | 480 |
|  | NAS Gateway e Management | FS interface, peer decoupling, RAID, backup automatico, recovery, management | € 10.000 | € 2.000 | € 12.000 | 5 | **€ 60.000** |  |  |  |  |  |

# Diagramma della architettura del singolo nodo



## Configurazione della Cloud

La cloud verrà realizzata mediante la installazione di X nodi all’interno della Regione Campania, Puglia e Sicilia, interconnessi logicamente in modo da fornire servizi unitari mediante la architettura software precedentemente presentata.

La configurazione totale del sistema, pertanto, sarà di:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Capacità di calcolo e storage** | | | | |
| **Nodo** | **Regione** | **Totale** | **CPU(#)** | **Core(#)** | **GPU(#)** | **RAM(GB)** | **Dischi (Gb)** |
| Nodo 1 | Campania | € 2.227.064 | 68 | 272 | 16 | 2560 | 535,2 |
| Nodo 2 | Sicilia | € 2.227.064 | 68 | 272 | 16 | 2560 | 535,2 |
| Nodo 3 | Puglia | € 2.227.064 | 68 | 272 | 16 | 2560 | 535,2 |
| ***Totale*** |  | ***€ 6.681.191*** | ***204*** | ***816*** | ***48*** | ***7680*** | ***1605,6*** |