Scalabilità Elastica e Multitenancy

Sistemi Concorrenti e Distribuiti

Lorenzo Bordoni

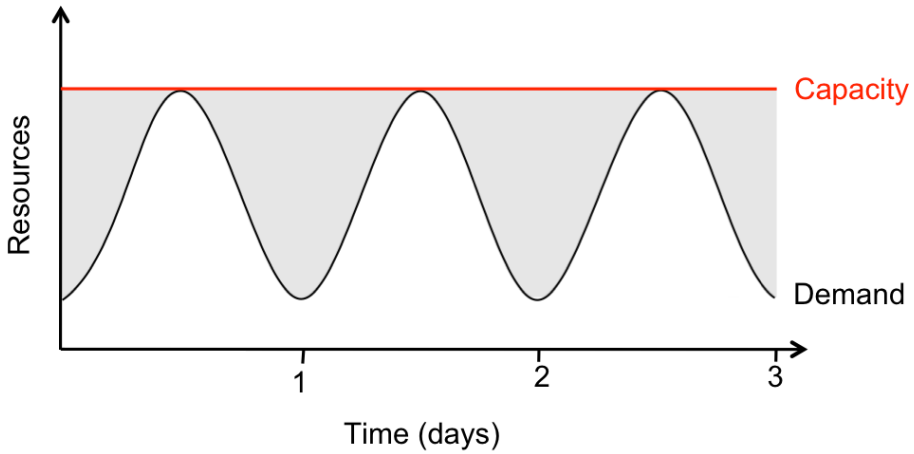
Corso di Laurea Magistrale in Informatica
Università di Padova

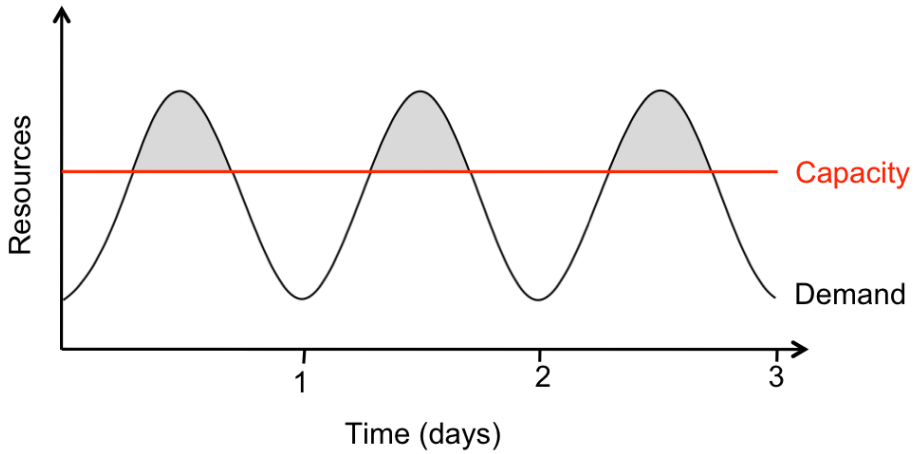
12 ottobre 2016

Elasticità Rapida

Rapid elasticity: [Computing, Storage, Networking] Capabilities elastically provisioned and reclaimed, automatically, to scale rapidly outward and inward to an extent commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time

—The NIST Definition of Cloud Computing

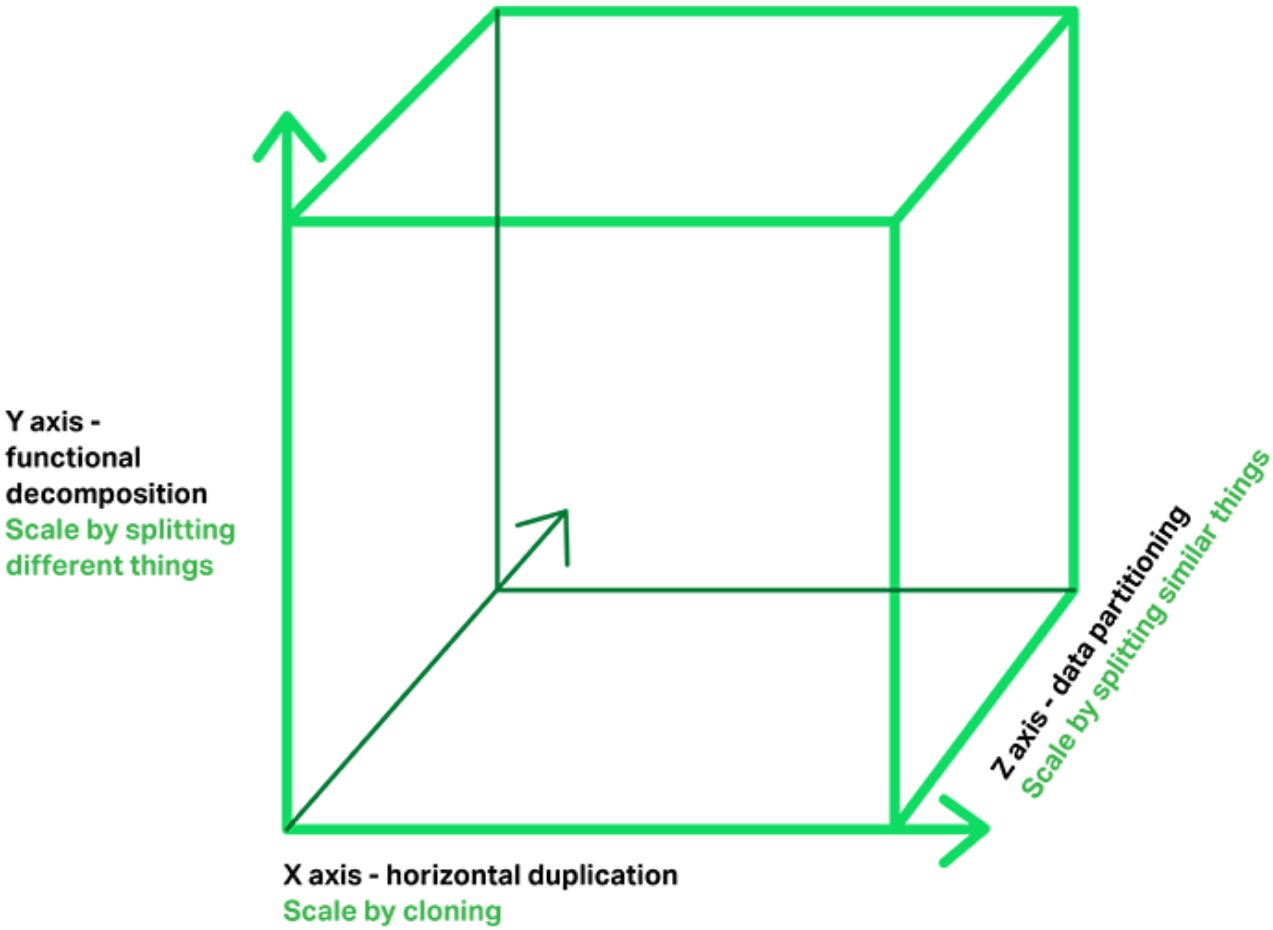




Lesson Learned

Nel Cloud, preallocare non è un'opzione

The Scale Cube



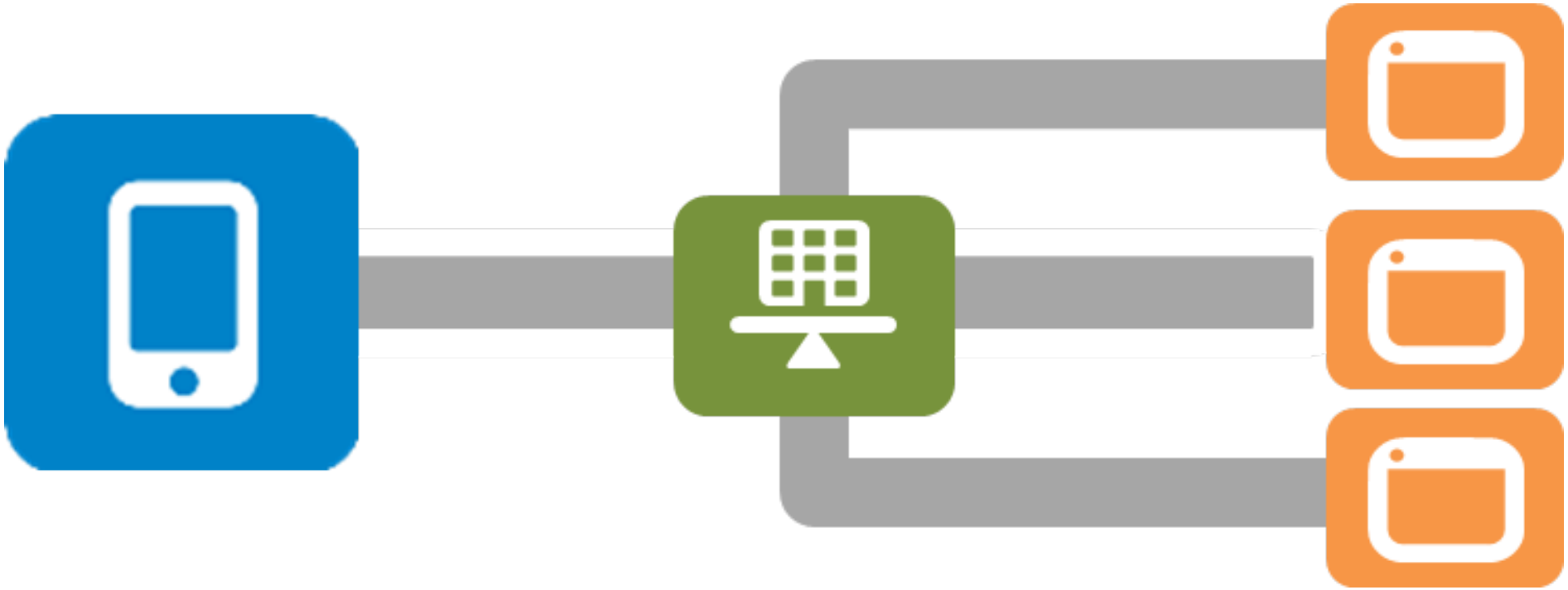
—The Art of Scalability

Scalabilità Verticale

Lesson Learned

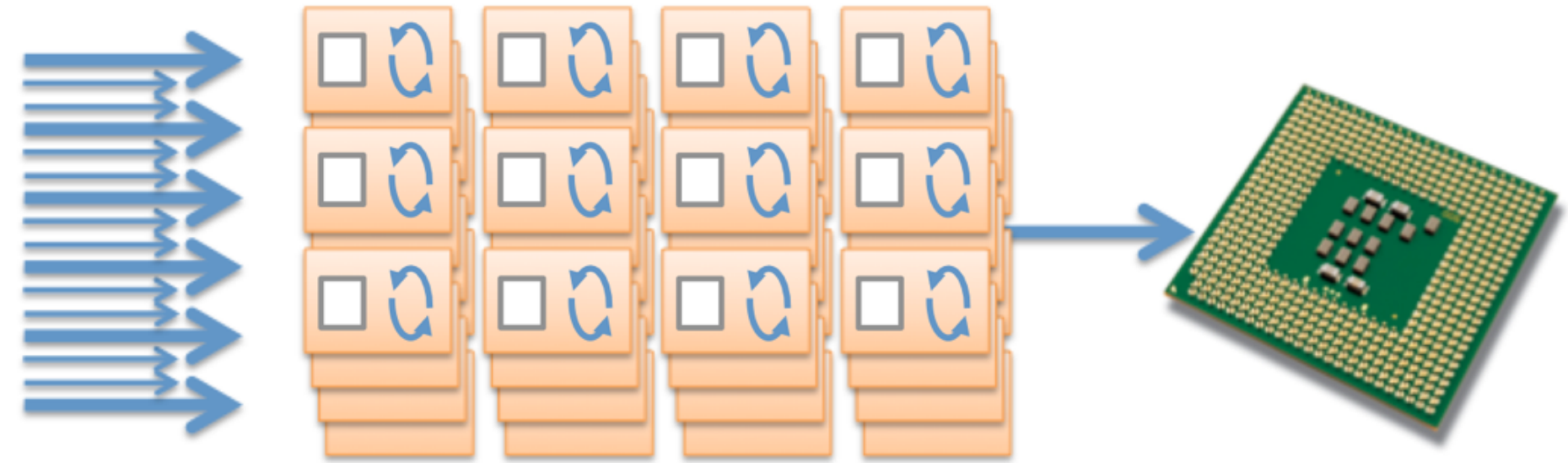
Scalare verticalmente usando HW più potente è una strategia destinata a fallire, poiché molte soluzioni Cloud tipicamente affrontano carichi che una singola macchina non può gestire, qualunque sia la sua capacità

Scalabilità sull'Asse X (*cloning*)



Replicare un'applicazione o un servizio affinché il lavoro possa essere facilmente distribuito tra le istanze

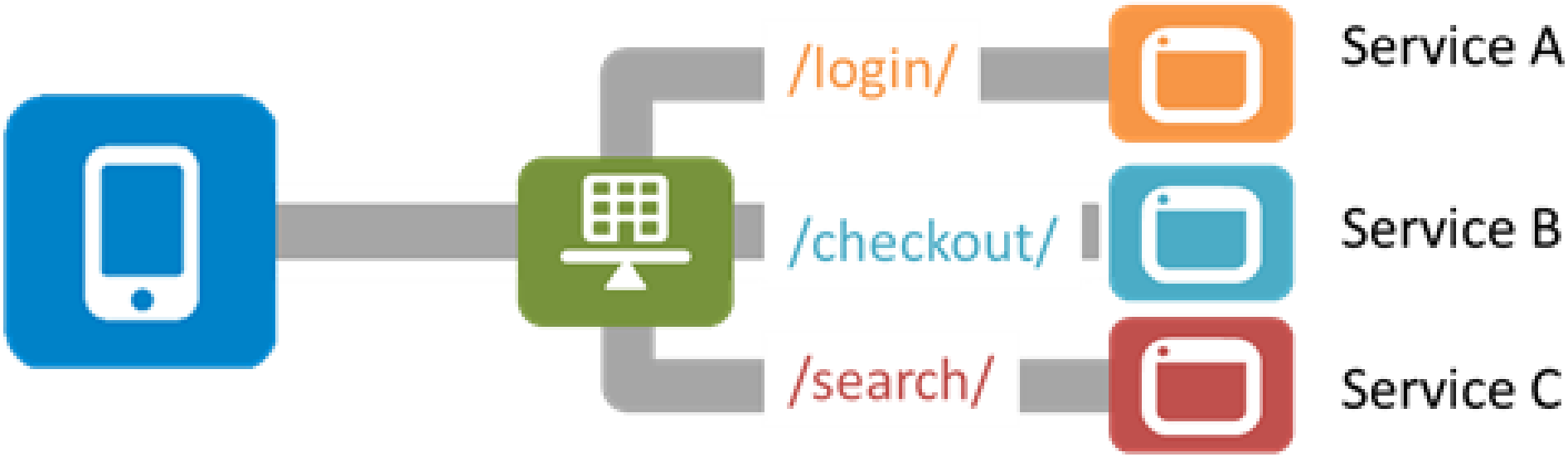
Scalabilità Orizzontale



Il limite di Apache HTTP Server

Un processo riceve le richieste e le ripartisce internamente ai suoi *worker*, che vivono nello spazio di memoria del padre. Il limite resta la capacità del singolo nodo ospite

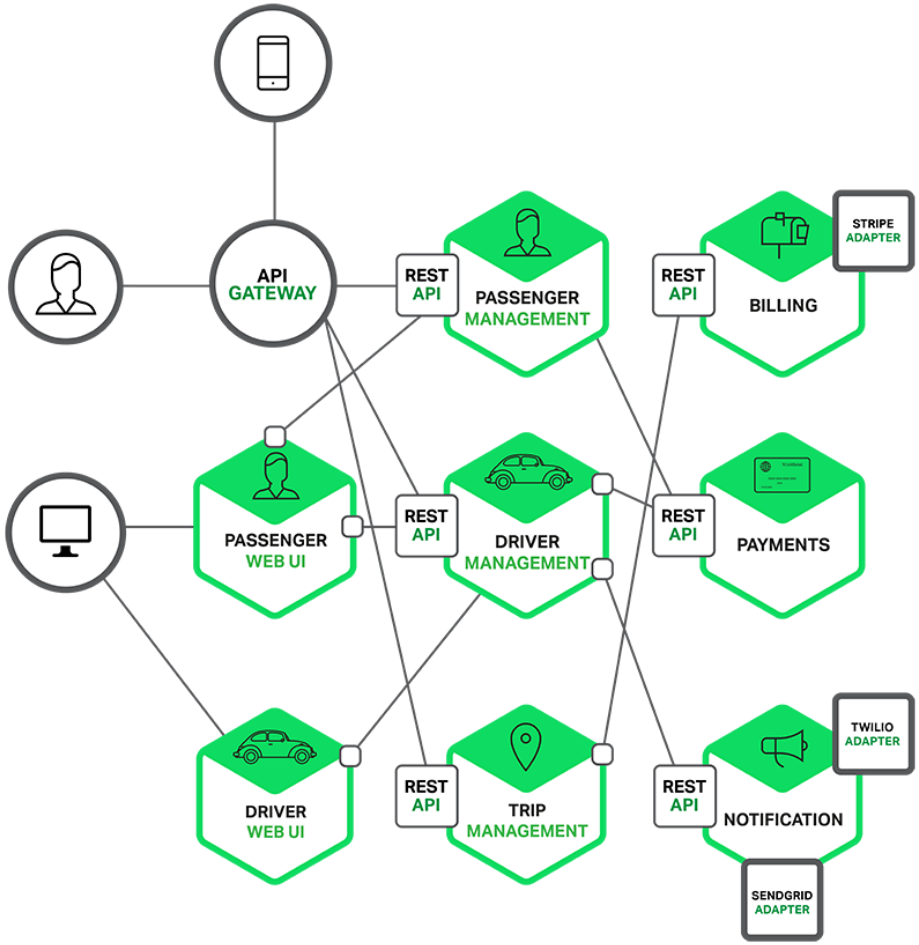
Scalabilità sull'Asse Y (*staging*)



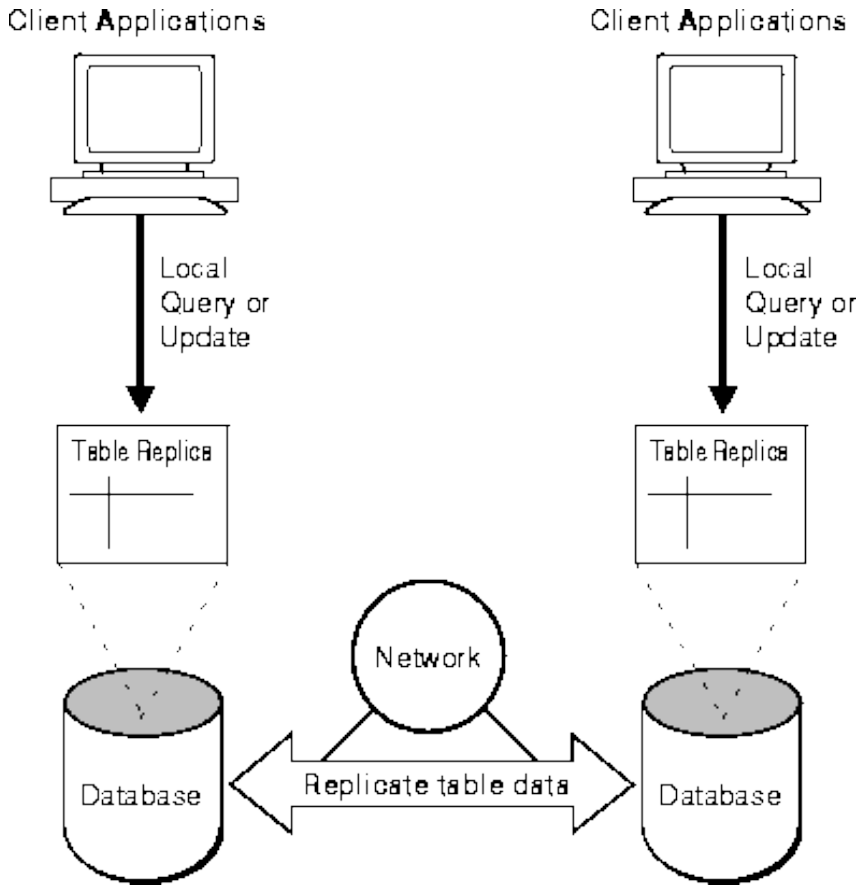
Ripartire il lavoro per tipo di servizio o di funzione all'interno dell'applicazione (*pipelining*, *staging*)

Disaccoppiamento dei Servizi

- *Service orientation*
- *Microservices*
- *Loose coupling*
- *Statelessness*



Replicazione dei Dati



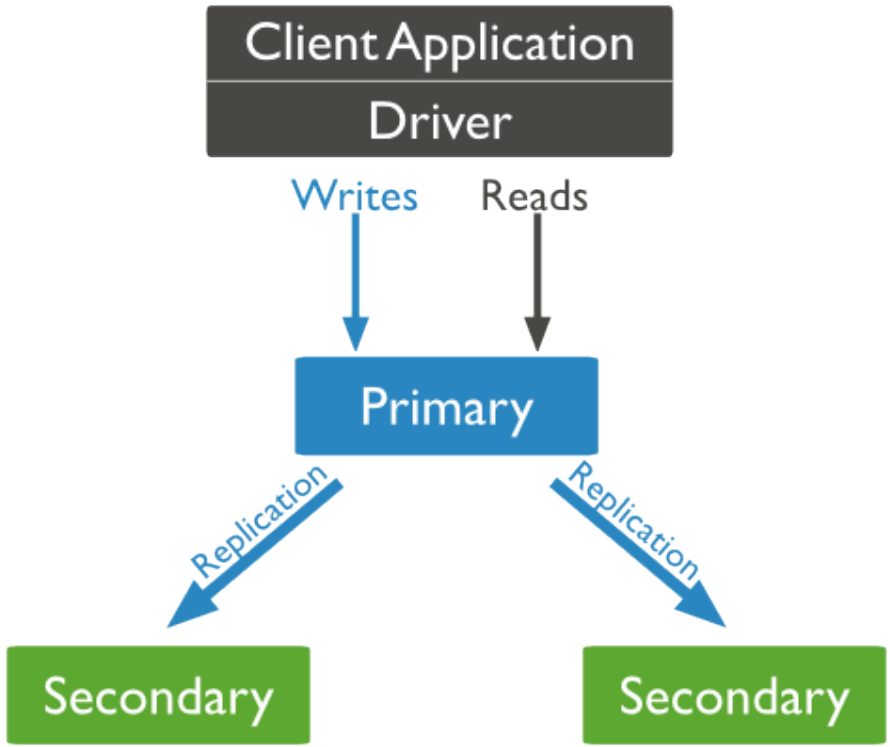
Lesson Learned

La replicazione di dati sotto transazioni compromette le prestazioni: più transazioni contemporanee, più conflitti, più ritardi

Replicazione in MongoDB

Eventual Consistency

Letture e scritture avvengono sulla replica primaria. L'applicazione può opzionalmente leggere da repliche secondarie, in cui i dati potrebbero non essere sempre aggiornati

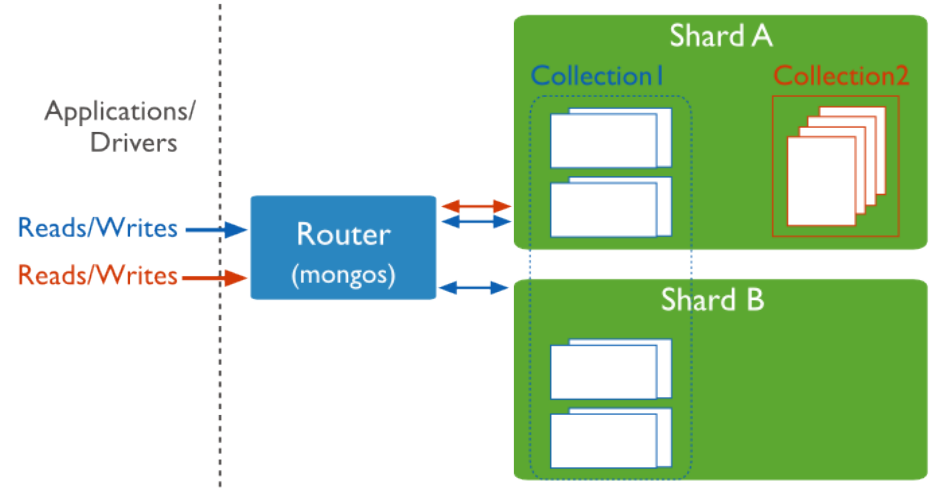


Scalabilità sull'Asse Z (*sharding*)

Dividere i dati secondo gli attributi, che sono cercati o determinati al momento dell'interrogazione

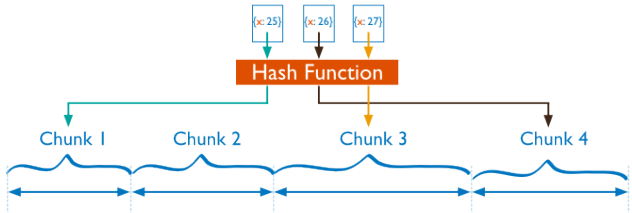
Sharding in MongoDB

Sharded cluster

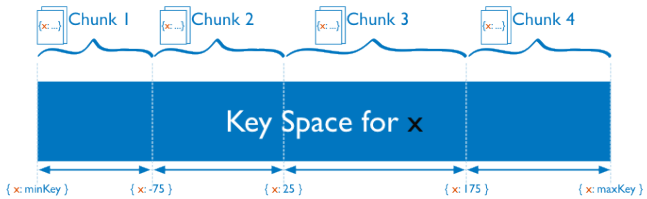


Strategie di sharding

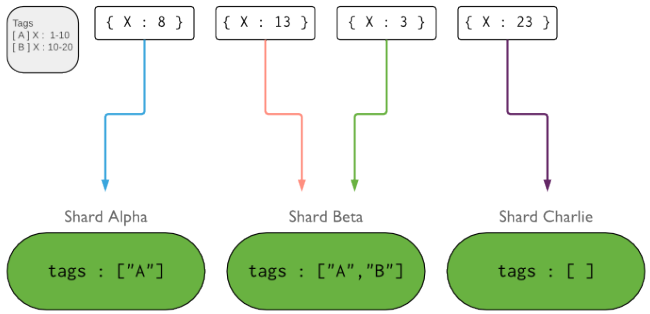
- Hashed sharding



- Ranged sharding



- Tag Aware sharding



Multitenancy

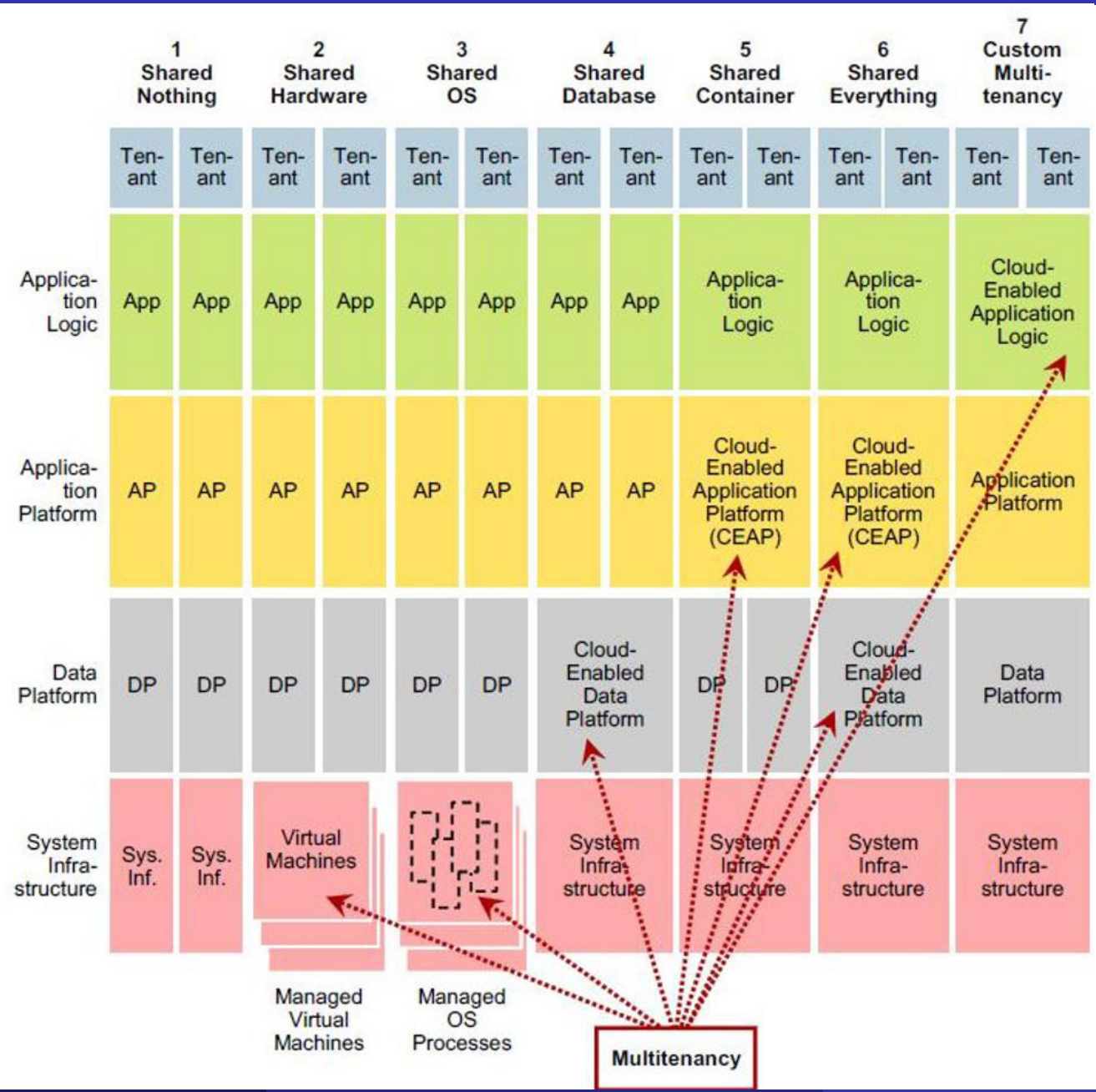
Resource pooling: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth

—The NIST Definition of Cloud Computing

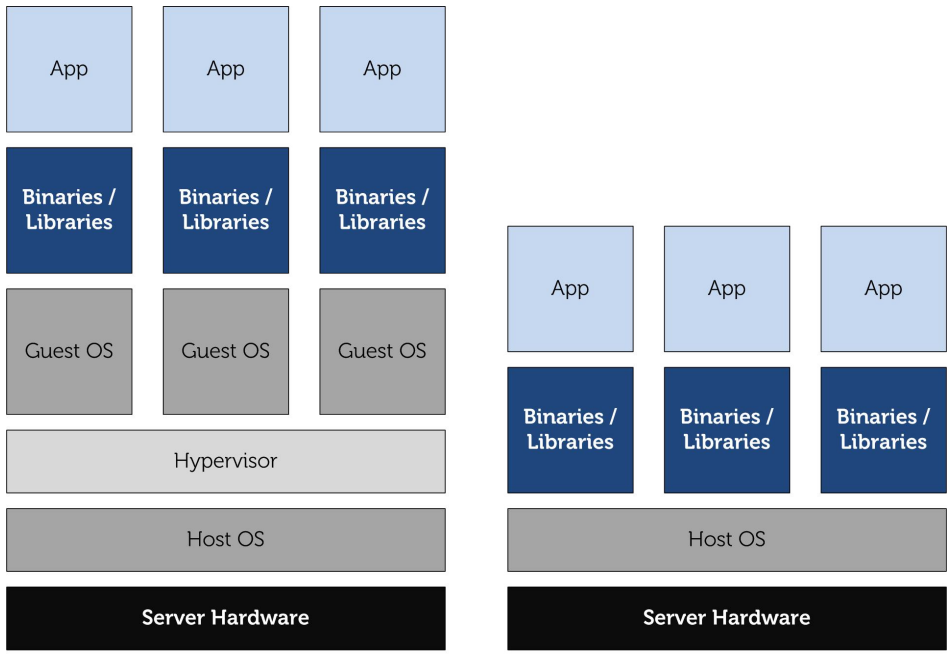
Obiettivo Principale

Ottenere un elevato utilizzo dell'HW, per ridurre i costi, senza compromettere le prestazioni o la sicurezza dei *tenant*

Gartner Reference Model for Multitenancy



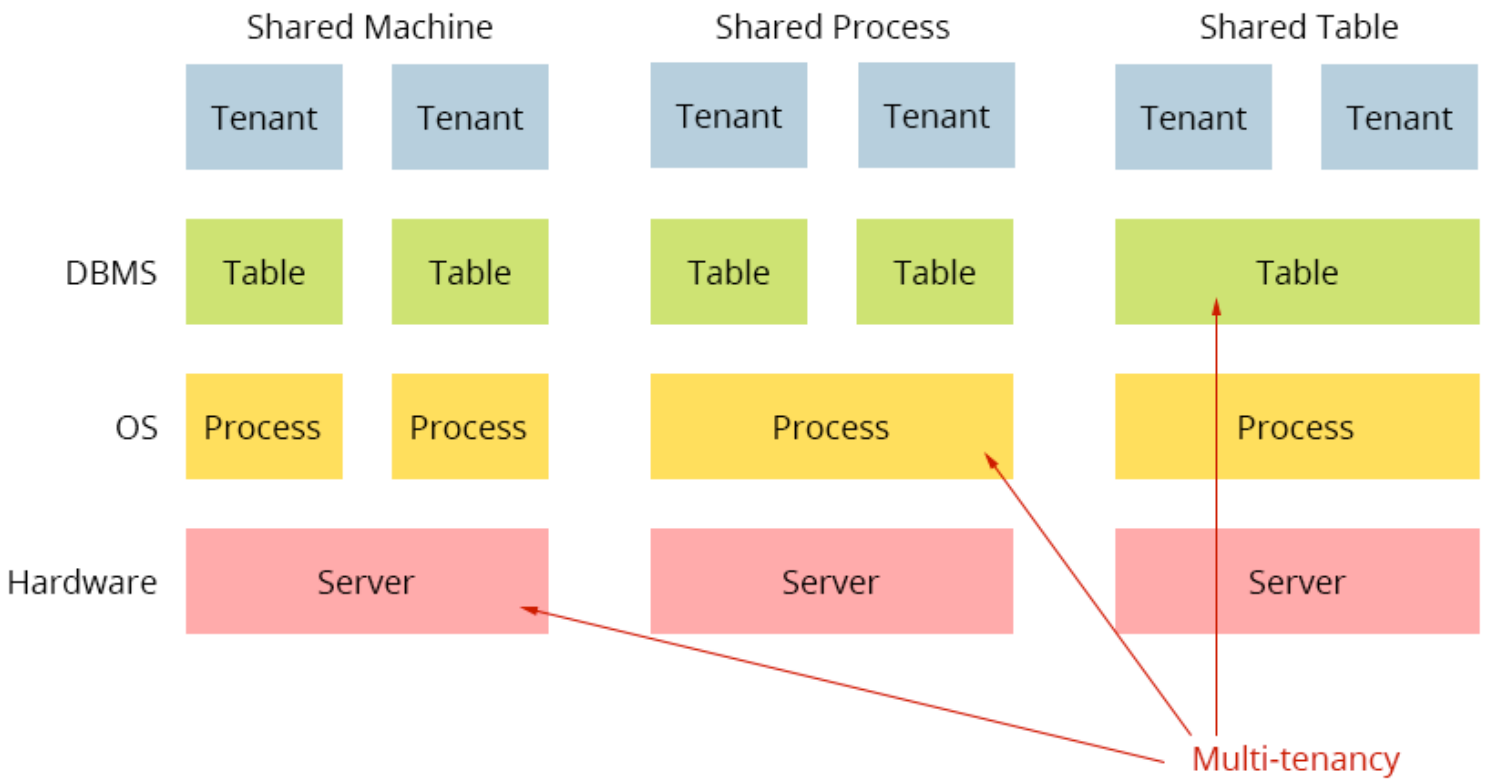
Multitenancy dei Servizi nel Cloud PaaS



Lesson Learned

I *container* e le macchine virtuali garantiscono un isolamento delle risorse simile, ma il diverso approccio architetturale permette ai *container* di essere più portabili ed efficienti

Multitenancy del DB nel Cloud PaaS



Lesson Learned

L'approccio *shared process* sembra il più promettente, ma attualmente la maggior parte dei DBMS non sono pronti a supportarlo. In genere, questo richiede di sfumare il confine tra dati e metadati