Cloud Computing

Definitions (I)

We have redefined Cloud Computing to include everything that we already do. I do not understand what we would do differently other then change the working of some of our ads.

Larry Ellison
Oracle CEO

It's stupidity. It's worse than stupidity: it's a marketing hype campaign. Somebody is saying this is inevitable – and whenever you hear somebody saying that, it's very likely to be a set of businesses campaigning to make it true.

Richard Stallman
GNU & FSF father

Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet).

Wikipedia

Definitions (III)

National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

Definitions (III)

Cloud computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the data centers that provide those services.

The services themselves have long been referred to as Software as a Service (SaaS), so we use that term.

The datacenter hardware and software is what we will call a Cloud.

When a Cloud is made available in a pay-as-you-go manner to the public [...] the service being sold is Utility Computing.

http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf

Properties and Characteristics

Scalability
Elasticity

- Dynamic provision
- Multi-tenant design

Manageability
Interoperability

- Autonomic Control
- System monitoring
- Billing System

Availability
Reliability

- Fault tolerance
- System resilience
- Security

Accessibility
Portability

- Uniform access
- Thin client

Performance
Optimization

- Parallel processing
- Load balancing
- Job scheduling

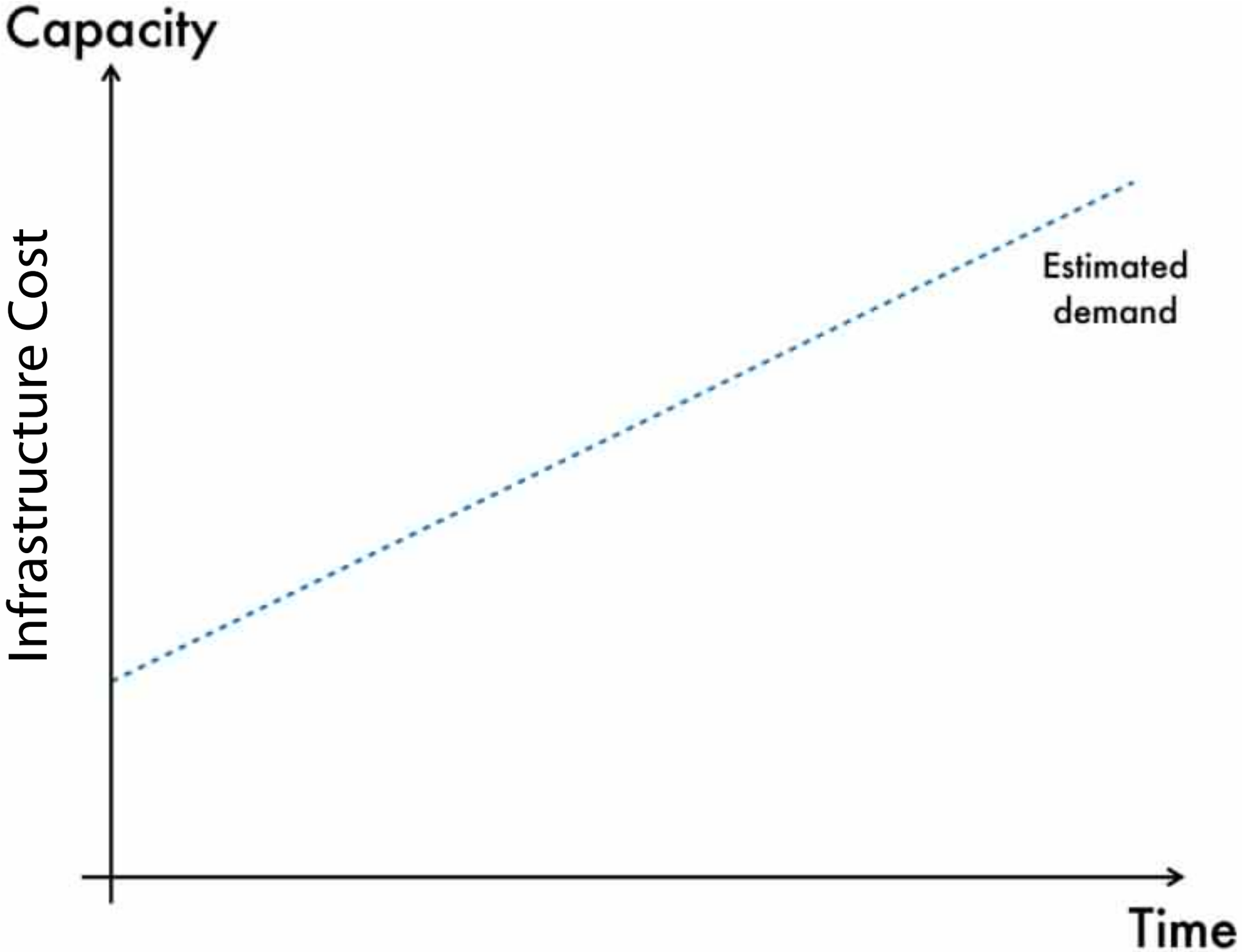
Scalability and Elasticity

- Scalability
  - A desirable property of a system, a network, or a process, which indicates its ability to either handle growing amounts of work in a graceful manner or to be readily enlarged.

- Elasticity
  - The ability to apply a quantifiable methodology that allows for the basis of an adaptive introspection with in a real time infrastructure.

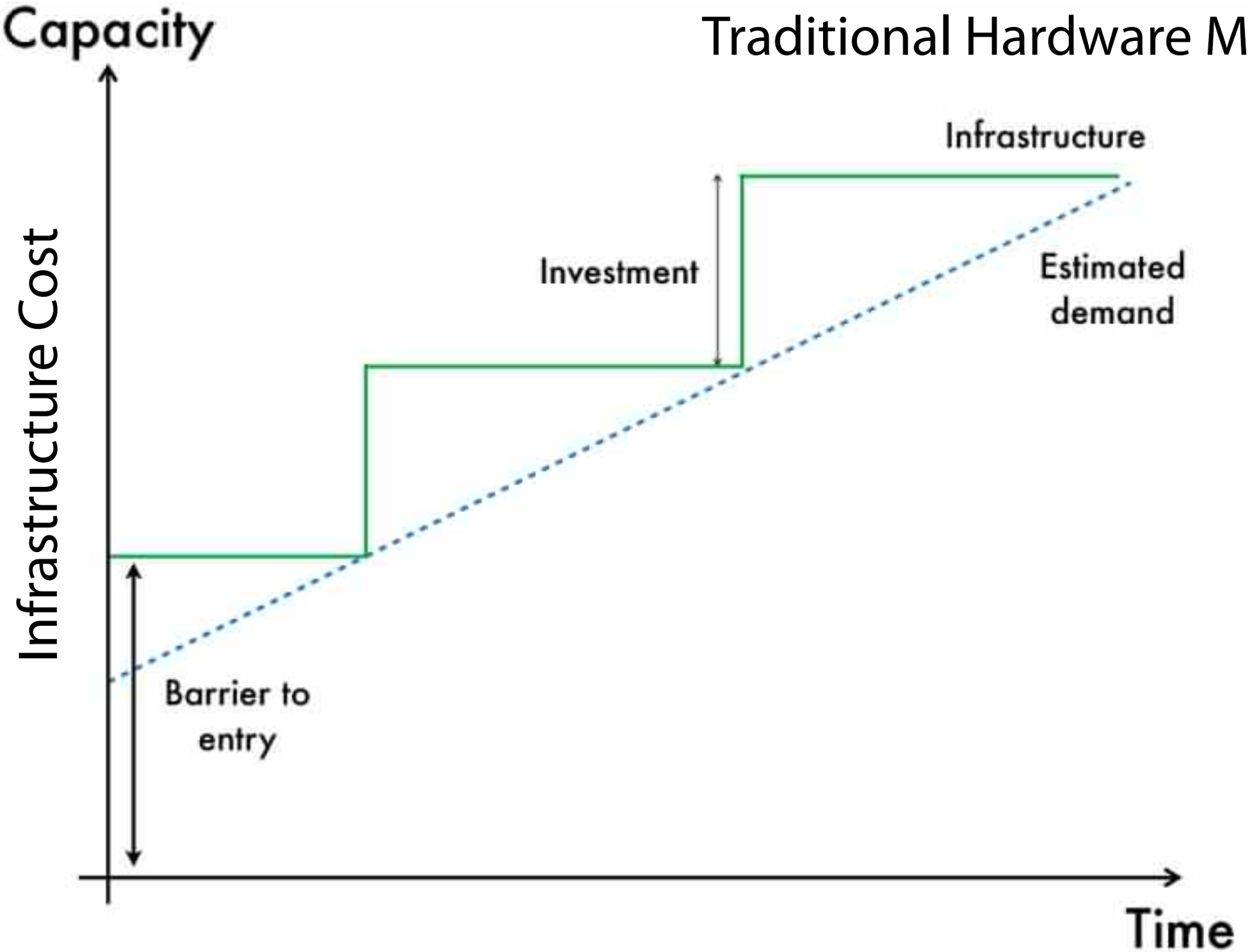
- How to achieve?
  - Dynamic provisioning
  - Multi-tenant design

Elasticity (I)

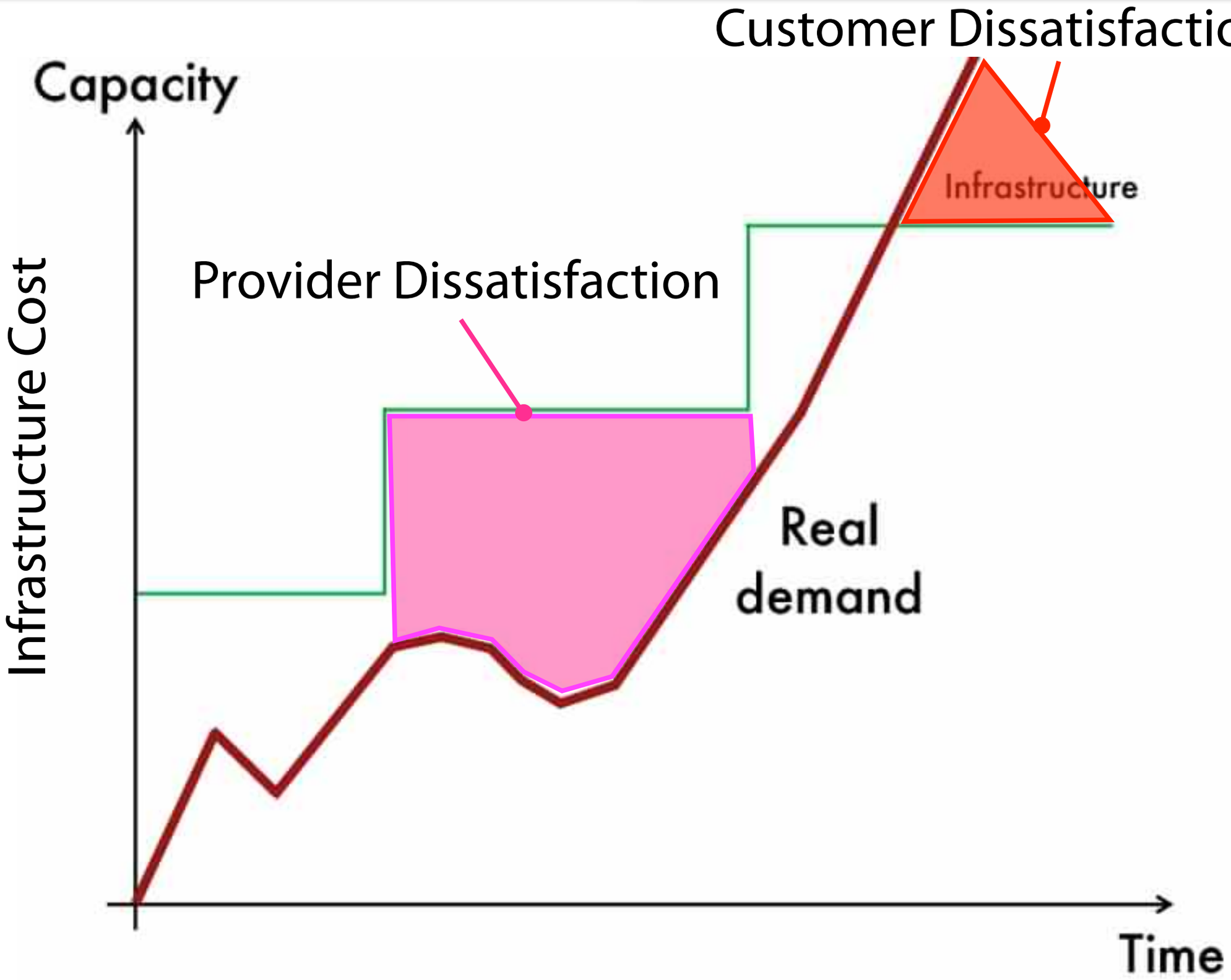


Elasticity (II)

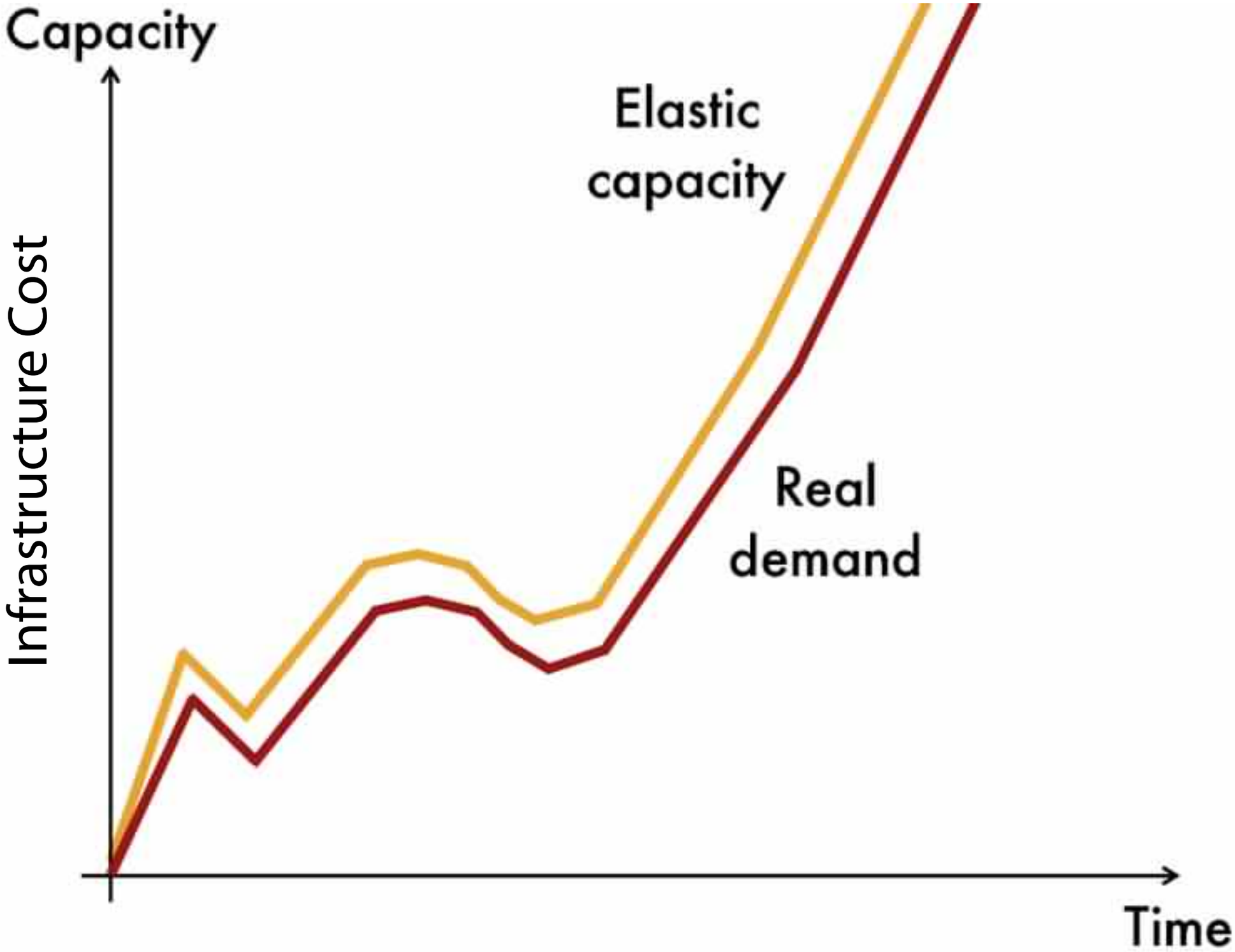
Traditional Hardware Model



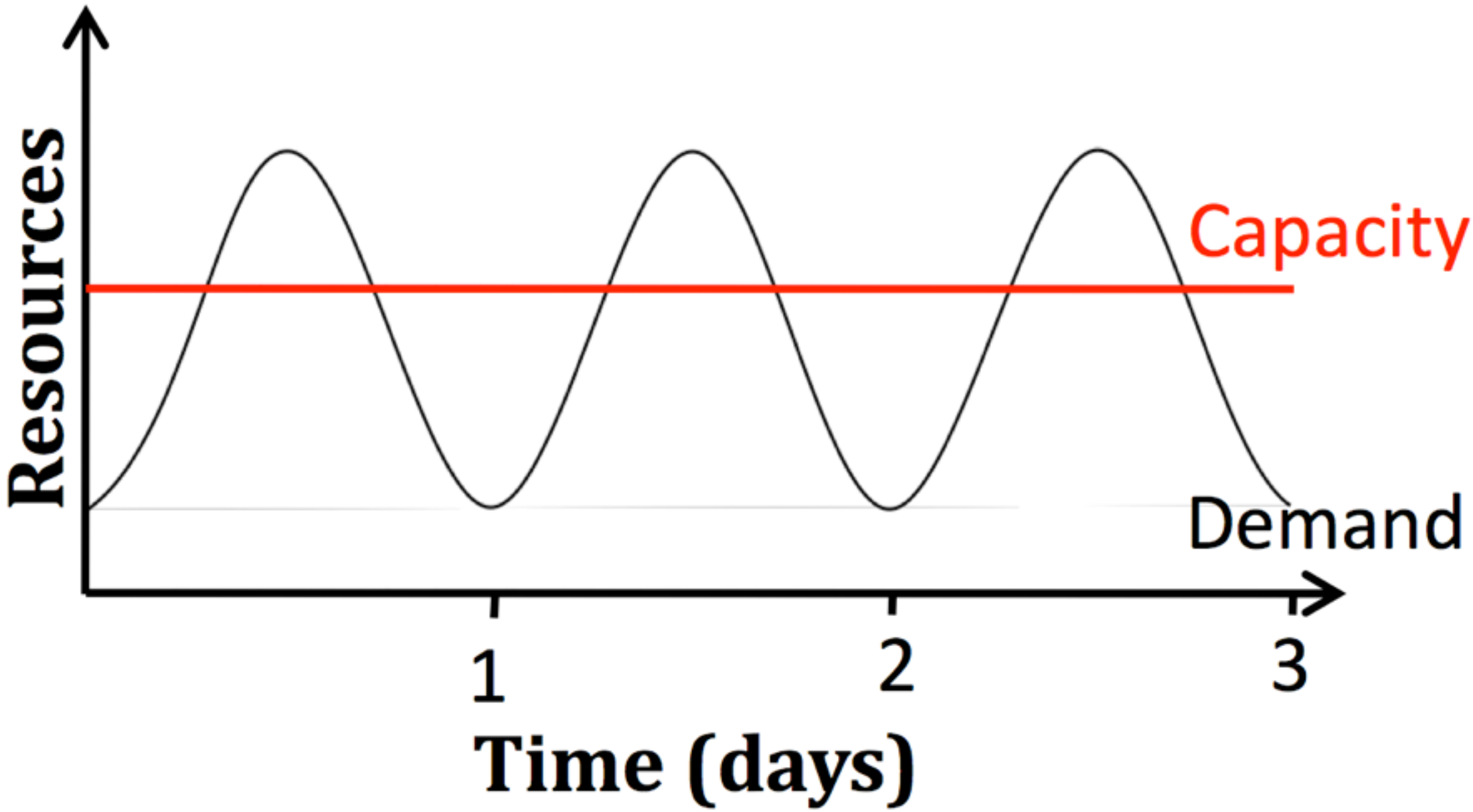
Elasticity (III)



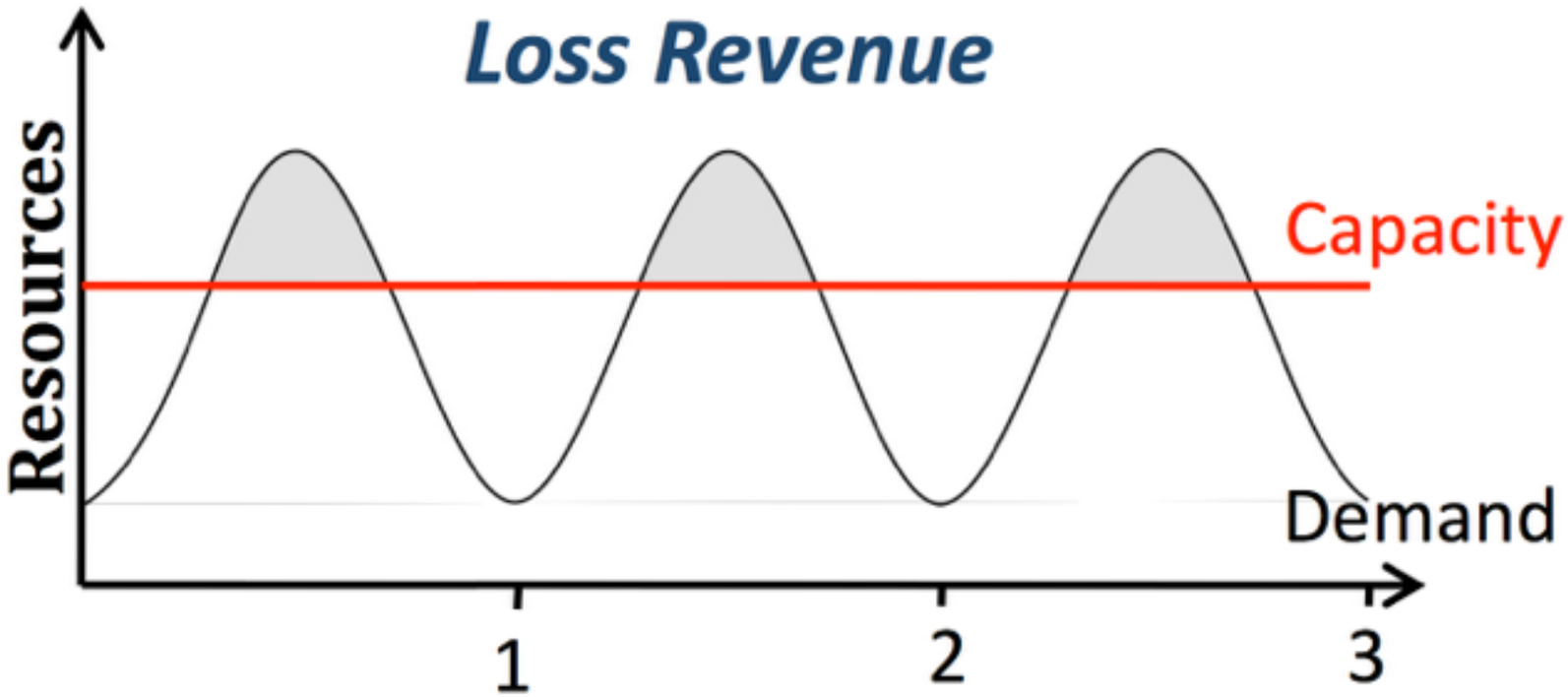
Elasticity (IV)

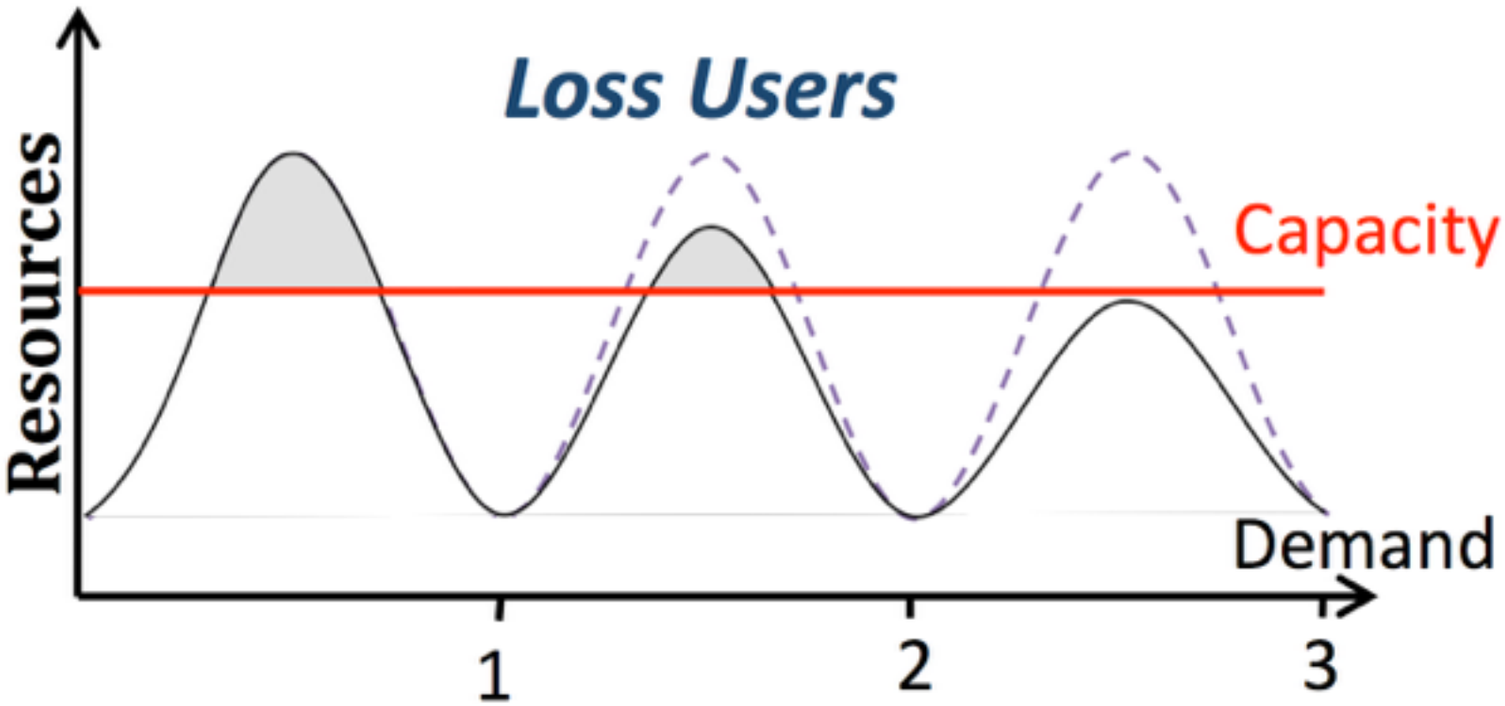


Traditional Provisioning

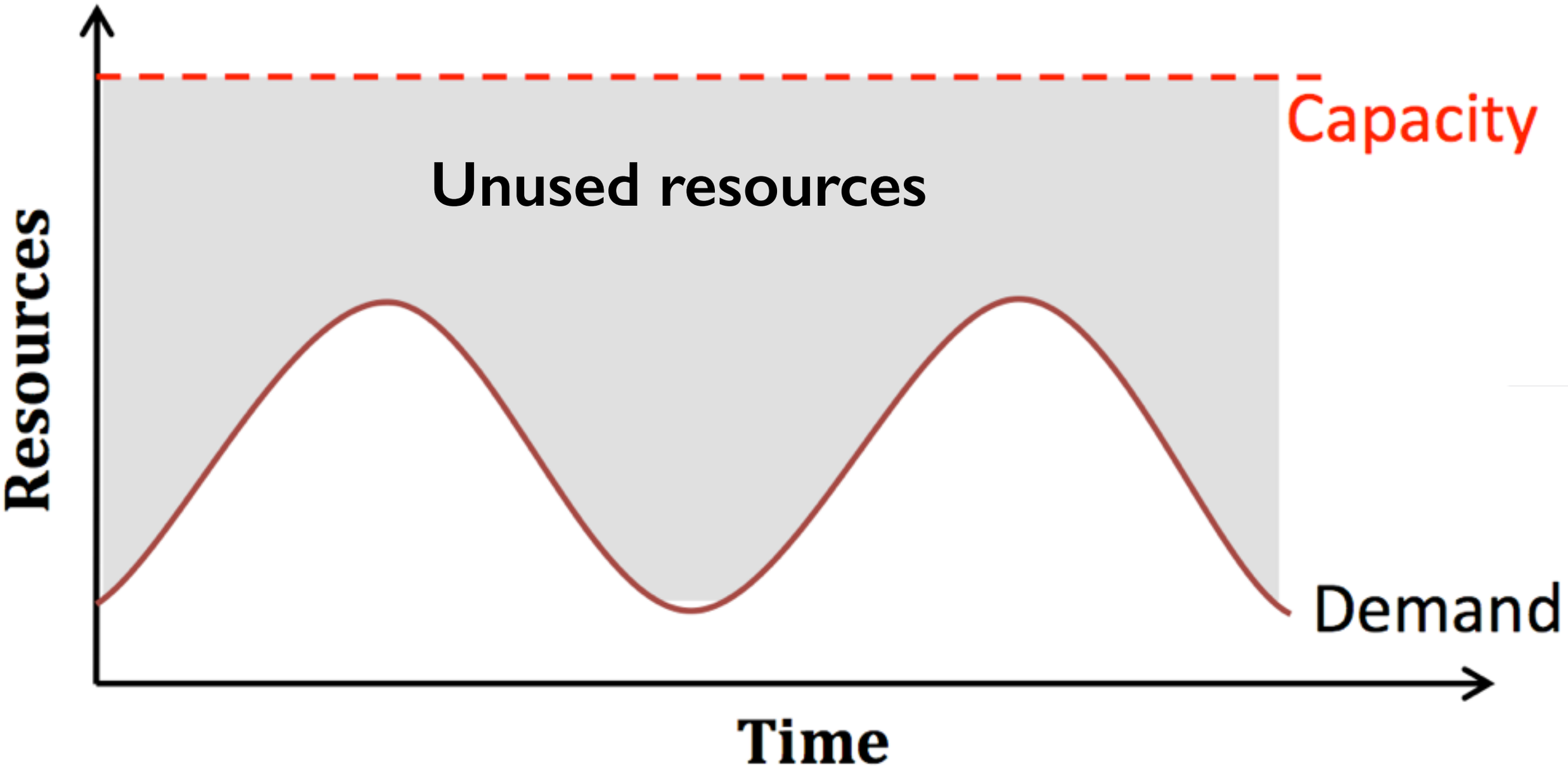


Under-provisioning



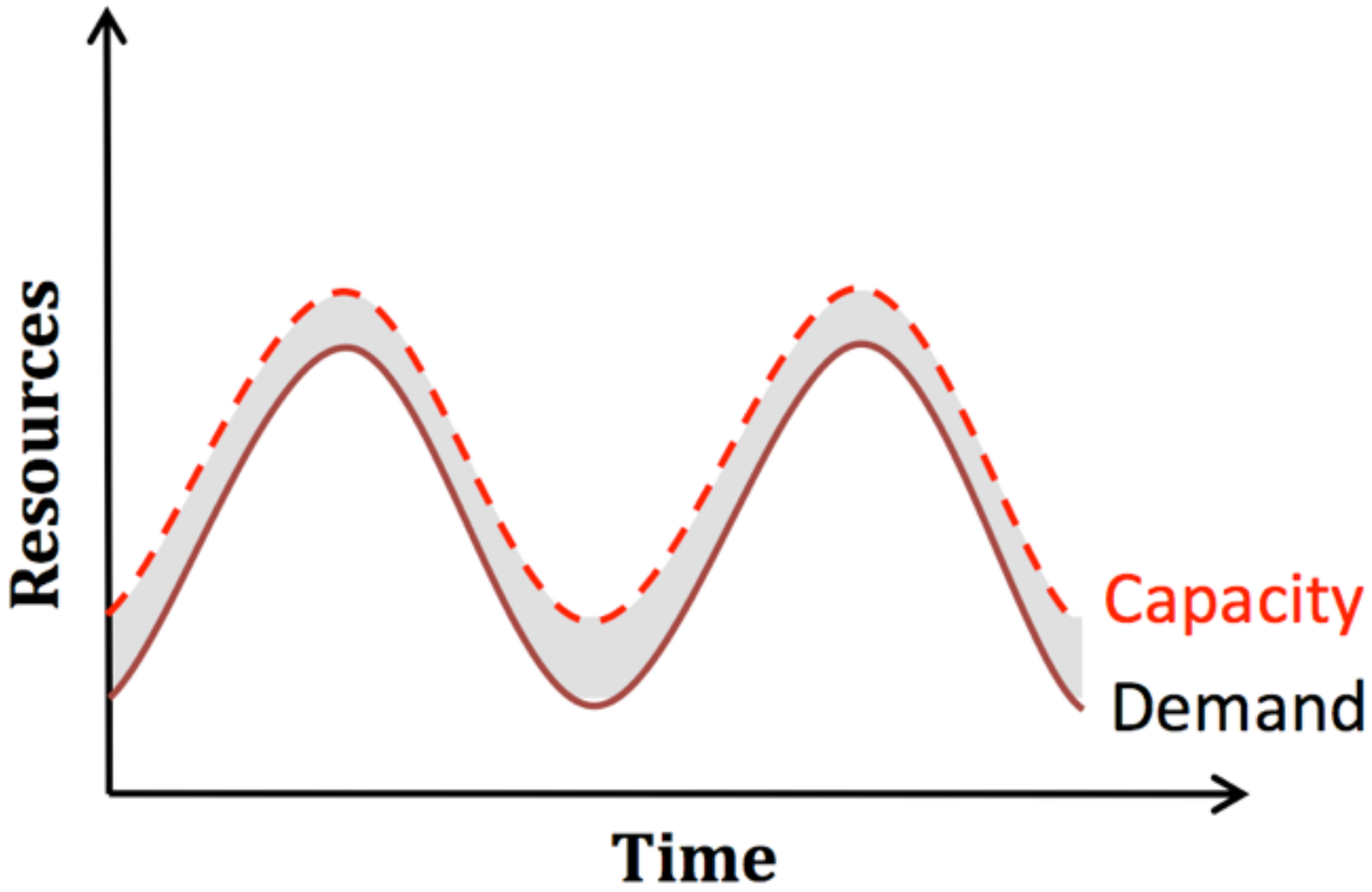


Over-provisioning



Dynamic Provisioning

- Cloud resources should be provisioned dynamically
  - Meet seasonal demand variations
  - Meet demand variations between different industries
  - Meet burst demand for some extraordinary events



Multi-tenant Design

- Multi-tenant refers to a principle in software architecture where a single instance of the software runs on a server, serving multiple client organizations.
  - Multi-tenancy is contrasted with a multi-instance architecture where separate software instances (or hardware systems) are set up for different client organizations
  - With a multi-tenant architecture, a software application is designed to virtually partition its data and configuration thus each client organization works with a customized virtual application instance.
- Client requirements:
  - Multi-tenant applications are typically required to provide a high degree of customization to support each target organization's needs.
  - Multi-tenant applications are expected to provide adequate levels of security and robustness.

Availability and Reliability

- Availability
  - The degree to which a system, subsystem, or equipment is in a specified operable and committable state at the start of a mission, when the mission is called for at an unknown time.
- Reliability
  - The ability of a system or component to perform its required functions under stated conditions for a specified period of time.

- How to achieve?
  - Fault tolerance
  - System resilience
  - Security

Fault Tolerance

- Fault-tolerance is the property that enables a system to continue operating properly in the event of the failure of some of its components.
- If its operating quality decreases at all, the decrease is proportional to the severity of the failure, as compared to a naively-designed system in which even a small failure can cause total breakdown.

- Four basic characteristics :
  - No single point of failure
  - Fault detection and isolation to the failing component
  - Fault containment to prevent propagation of the failure
  - Availability of reversion modes

System Resiliency

- Resilience is the ability to provide and maintain an acceptable level of service in the face of faults and challenges to normal operation.

- Resiliency pertains to the system's ability to return to its original state after encountering trouble. In other words, if a risk event knocks a system offline, a highly resilient system will return back to work and function as planned as soon as possible.

- Disaster recovery is the process, policies and procedures related to preparing for recovery or continuation of technology infrastructure critical to an organization after a natural or human-induced disaster.

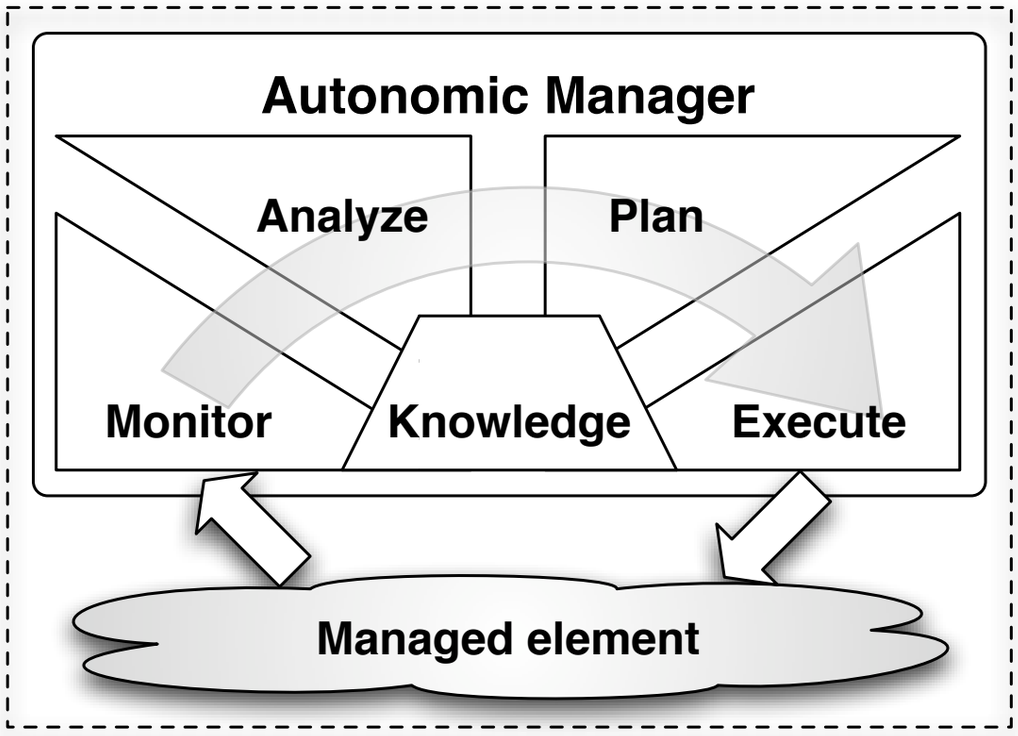
Security

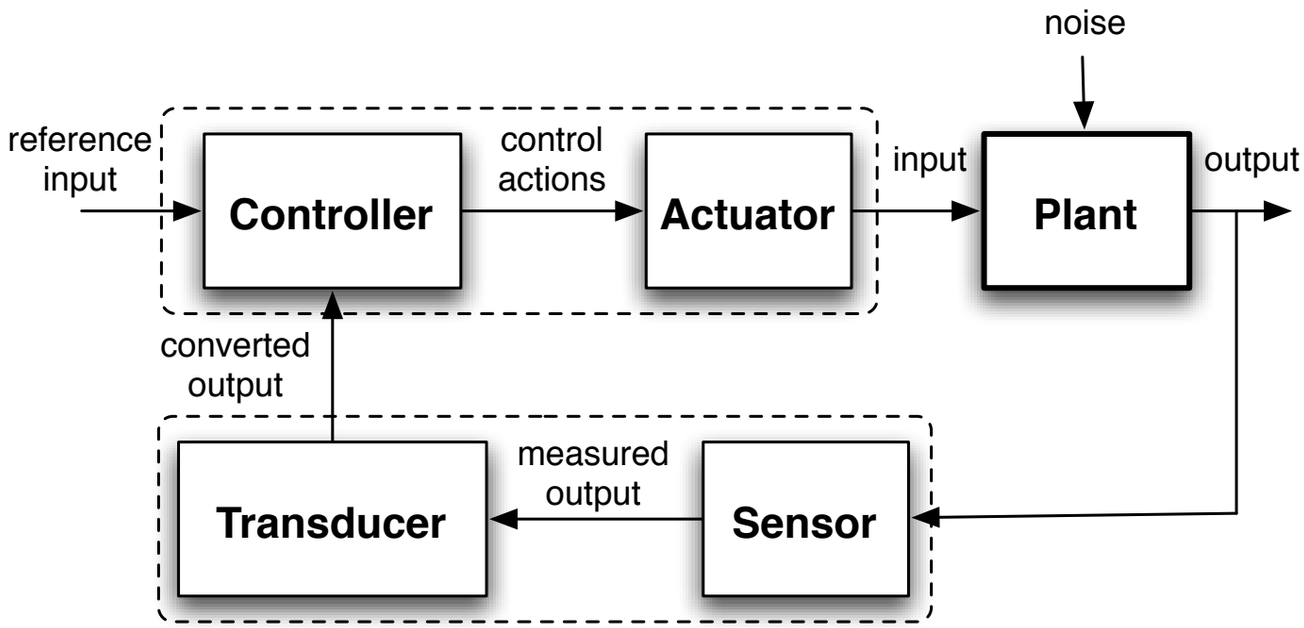
- Cloud security is an evolving sub-domain of computer security, network security, and, more broadly, information security.
- It refers to a broad set of policies, technologies, and controls deployed to protect data, applications, and the associated infrastructure of cloud computing.
- Important security and privacy issues:
  - Data Protection
    - To be considered protected, data from one customer must be properly segregated from that of another.
  - Identity Management
    - Every enterprise will have its own identity management system to control access to information and computing resources.
  - Application Security
    - Cloud providers should ensure that applications available as a service via the cloud are secure.
  - Privacy
    - Providers ensure that all critical data are masked and that only authorized users have access to data in its entirety.

Manageability and Interoperability

- Manageability
  - Enterprise-wide administration of cloud computing systems.
  - Systems manageability is strongly influenced by network management initiatives in telecommunications.
- Interoperability
  - Interoperability is a property of a product or system, whose interfaces are completely understood, to work with other products or systems, present or future, without any restricted access or implementation.
- How to achieve?
  - Autonomic Control
  - System Monitoring
  - Billing System

Autonomic Control





- Autonomic Computing
  - Its ultimate aim is to develop computer systems capable of self-management, to overcome the rapidly growing complexity of computing systems management, and to reduce the barrier that complexity poses to further growth.