The three most common strategies are:

• **All eggs in one basket:** One machine used for many purposes

• **Beautiful snowflakes:** Many machines, each uniquely configured

• **Buy in bulk, allocate fractions:** Large machines partitioned into many smaller virtual machines using virtualization or containers

In addition, there are variations and alternatives:

• **Grid computing:** Many machines managed one as unit

• **Blade servers:** A hardware architecture that places many machines in one chassis

• **Cloud-based compute services:** Renting use of someone else’s servers

• **Software as a service (SaaS):** Web-hosted applications

• **Server appliances:** Purpose-built devices, each providing a different service

The phrase **“web server”** might refer to a host being used to provide a web site (a machine) or the software that implements the HTTP protocol (Apache HTTPD).

Unless specified, this book uses the term **“server”** to mean a machine. It refers to a **“service”** as the entire hardware/software combination that provides the service users receive.

The term **forklift upgrade** is industry slang for a wholesale replacement. In such a situation you are removing one machine with a metaphorical forklift and dropping a replacement in its place.

This Catch-22 is known as **dependency hell.**

The excess capacity is called **stranded capacity** because it is unusable in its current form.

continuous integration (CI)

total cost of ownership (TCO)

continuous integration/continuous deployment (CI/CD)

Software as a service (SaaS)

A blade server has many individual slots that take motherboards, called **blades**, that contain either a computer or storage.

There are three common definitions for the cloud, each coming from different communities:

• Consumers • Business people • IT professionals

Some of the characteristics that differ for servers versus workstations are profiled here:

• More CPU performance • High-performance I/O • Expandability • Upgrade options • Rack mountable • Front and rear access • High-availability options • Remote management

out-of-band (OOB)

Integrated Lights-Out (iLO)

Intelligent Platform Management Interface (IPMI)

RAID (Redundant Array of Independent Disks)

lights-out management (LOM)

remote insight board (RIB)

Remote Insight Light-Out Edition (RILOE)

power distribution unit (PDU)

service level agreements (SLAs)

The term **N+ 1 redundancy** is used when we wish to indicate that there is enough spare capacity for one failure. **N+ 2 redundancy** would mean there is enough spare capacity for two failed power supplies.

**Hot-swap** refers to the ability to add, remove, and replace a component while the system is running.

Be mindful of components that are labeled **hot-plug.** This means that it is electrically safe for the part to be replaced while the system is running, but the part may not be recognized until the next reboot.

A **KVM switch** is a device that lets many machines share a single keyboard, video screen, and mouse (KVM).

The spares kit is less expensive than buying a second machine, often called a **cold spare** because it is left powered off (cold) until needed.

This practice is called **cross-shipping**; the parts cross paths as they are delivered.

Since each program is sharing the CPU, each program runs for a few milliseconds, and then the OS puts it to sleep and runs another program. Each time one process is put to sleep so that another can run is called a **context switch.**

virtual machine manager (VMM)

graphics processing units (GPUs)

Non-Uniform Memory Architecture (NUMA)

Silicon Graphics (SGI)

network interface cards (NICs)

RAID (redundant array of independent disks)

Allocating more resources than exist is called **oversubscription.**