QoS

Giving priority to certain network traffic to minimize delay and packet loss.
POE: allows devices to receive power over an Ethernet cable.
Traditional phones operate over the public switched telephone network (PSTN) /(POTS (Plain Old telephone service)).
IP phones use VoIP (Voice Over IP) to enable phone calls over IP network as internet, they are connected to a switch like endhosts, they have an internal 3-port switch, one is uplink to external switch, one is downlink to pc and one is internally connected to the phone allowing the phone and the pc to share the same switch port.
It's recommended to separate voice traffic and data traffic by placing them in different VLANs by configuring a voice VLAN, so traffic from pc is untagged while from phone will be tagged with a VLAN id.
Using **command: switchport voice vlan <vlan number>,** sw1 will use CDP to tell phone to tag its traffic in vlan 11**.**
Power Source Equipment (PSE) as a switch can power (Powered Devices) (PD) as IP phone over an Ethernet cable using POE.
When a device is connected to a PoE enabled port, the PSE sends low power signals, monitors the response and determines how much power the PD needs, if the device needs power, the PSE supplies the power to allow PD to boot and continues to monitor the PD and supply the required amount of power.
Power Policing: can be configured to prevent a PD from taking too much power.
Command: power inline police / power inline police action err-disable disables the port and send a syslog message if a PD draws too much power, interface is in error-disabled state and can be re-enabled with **shut** followed by **no shut**.
Command: power inline police action log: doesn't shutdown the interface if a PD draws too much power, it restarts the interface, sends a syslog message, the connected PD restarts and re negotiate its power needs.
Command: show power inline police <interface>.

Voice traffic used PSTN, data traffic used ip network as enterprise WAN, internet, ...etc, in modern networks all share the same IP network, enabling cost savings and integrations with collaboration software, the different kind of traffic have to compete for bandwidth, Qos is a set of tools used by network devices to apply different treatment to different packets.
QoS: manages → 1- Bandwidth: it allows to reserve a certain amount of link's bandwidth for specific kind of traffic.
2- Delay: one-way or two-way delays. 3- Jitter: the variation in one way delay between packets sent by the same app, IP phones have a jitter buffer to provide a fixed delay to audio packets.
4- Loss: the % of packets don't reach the destination caused by faulty cables or when a device's packet queue get full and the device starts discarding packets.
For acceptable interactive audio quality:
One way delay: 150ms or less.
Jitter: 30 ms or less. Loss: 1% or less.
Queuing: if a network device receives messages faster than it can forward them out of appropriate interface, the messages are queued, if queue is full new packets will be dropped (tail drop).
TCP global synchronization: network congestion→ tail drop→ global TCP window size decrease→ network underutilized → global TCP window size increase→ loop.
Randomly early detection (RED): when the amount of traffic in the queue reaches a certain threshold the device will start randomly dropping packets from select TCP flows.
Weighed Random Early Detection (WRED): allows you to control which packets are dropped depending on the traffic class at certain various thresholds.

Classification: organizes network traffic into classes using many methods as:
- an ACL: traffic permitted by an ACL is given a special treatment.
- NBAR (Network based Application Program): performs a deep packet inspection looking up to layer 7 to identify the specific kind of traffic.
In layer2 : PCP (CoS → class of service) 3-bits of dot1q tag is used to identify high/low priority traffic.
In layer3 : DSCP field of the IP header can also be used to identify high/low priority traffic.
PCP: 0 → Best effort (regular traffic). Call signaling traffic is marked by IP phones as PCP3 (critical apps) and voice traffic is marked as PCP5 →voice, can be used with trunk links for traffic not in native VLAN and access links with voice VLAN.
In IPV4: ToS (type of Service) byte: consists of DSCP and ECN, inold ToS three bits were used for IPP (IP precedence) and the remaining 5 bits were defined for various purposes mostly unused.
IPP: 6&7 for network control traffic as OSPF messages between routers, 5 for voice, 4 for video, 3 for voice signaling and 0 for best effort.
Modern: 6 bits for DSCP and 2 for ECN→ Default Forwarding (DF): best effort traffic, Expedited Forwarding (EF): low loss/latency/jitter as voice traffic, Assured Forwarding (AF) 12 values, Class Selector (CS) 8 values with backward compatibility with IPP (CS0 = DF) .
AFXY: X→ four traffic classes for priority forwarded with better service 3-bits. Y→ drop precedence (packets more likely to be dropped during congestion) 2-bits, the last bit is always 0.
To get DSCP: 8X+2Y.
CS: the extra 3-bits than IPP are removed giving 8 values. Voice traffic → EF, interactive video → AF4x, streaming video→ AF3x, high priority data→ AF2x, best effort ->DF.

Trust boundary: defines where devices trust/don't trust the QoS markings of received messages, if markings aren't trusted the device will change the markings according to the configured policy before forwarding them.
QoS uses multiple queues, the device matches traffic based on various factors for example DSCP and then place it in the appropriate queue, however the device is only able to forward one frame out of an interface at once so a scheduler is used.
Weighed round robin: packets are taken from each queue in order cyclically where more data is taken from higher priority queues.
CBWFQ (Class based weighed fair queuing): using a weighed round robin queue while guaranteeing a certain percentage of the interface's bandwidth during congestion.
Not ideal for voice/video because round robin can add delay/jitter as high priority traffic should wait for their turn in the scheduler.
Low Latency Queue (LLQ); designates one or more queues as strict priority queues, the scheduler takes the next packet from them until they are empty, can cause starvation to the other queues.
Within each queue congestion prevention tools as RED or WRED can be used.
Shaping: buffers traffic in queue if its rate goes over the configured rate (even if not maximum interface capacity).
Policing: drops the traffic if the traffic rate goes over the configured rate, burst traffic over the configured rate is allowed for a short period of time, the amount of burst traffic is configurable.
Classification can be used to allow for different rates for different kinds of traffic.
Per hop behavior: how each router prioritizes traffic over the next hop depending on its configuration.

Command: class-map <name>. Command: match protocol <protocol>.
Command: policy-map <name>. Command: set ip dscp <value>. Command: priority percent <percent>.
Command: bandwidth percent <percent>.
Command: service-policy output <name of policy>.