Performance Management

outline

- performance management def
- Purpose of performance management
- performance monitoring

- Businesses realize that their performance depend on the network.
- Network downtime has become a business issue instead of just a minor problem.
- Notions such as service level agreements (SLA) are imposed on the network to support specific business application requirements.
- There is expectation for connectivity to be available anytime, anywhere.

- this is impossible without intelligent systems managing the network.
- Need for technologies, processes, and applications in the area of Network Management

- The FCAPS model is an international standard defined by the International Telecommunication Union (ITU) that describes the various network management areas.
  - Fault
  - Configuration
  - Accounting
  - Performance
  - security

- performance metrics:
  - availability,
  - response time,
  - network round-trip time,
  - latency,
  - jitter,
  - reordering,
  - packet loss, etc.

Defining Performance Management

- Function: to evaluate and report upon the behaviour and effectiveness the network or network element.

- Role: to gather and analyze statistical data for purposes of monitoring and correcting the behaviour and effectiveness of the
  - network,
  - network element, or
  - other equipment and

  to aid in planning, provisioning, maintenance and the measurement of quality.

performance monitoring vs performance management.

- Performance monitoring collects
  - Device related,
  - Network related, and
  - Service related

  parameters and reports them via a graphical user interface, log files, etc.
- Performance management builds on the data collections but goes one step further by
  - actively notifying the administrator and
  - reconfiguring the devices if necessary
  - take other appropriate action

- performance monitoring vs performance management.
- An example is the data collection for SLAs.
- Performance monitoring would only collect the data and store it at a collection point.
- Performance management would
  - analyze the data and
  - compare against predefined thresholds and service definitions.
  - In the case of a service level violation, it then would generate a trouble ticket to a fault application or reconfigure the device.
  - For example, it would filter best-effort traffic or increase the committed access rate, and so on.

- performance management describes the following processes:
  - Performance monitoring
  - Data analysis
  - Performance management

1. Performance monitoring — Collecting network activities at the device level for the sake of

- Device-related performance monitoring
- Network performance monitoring
- Service performance monitoring

- Monitoring subtasks include
  - Availability monitoring
  - Response time reporting
  - Monitoring utilization (link, device, CPU, network, service, and so on)
  - Ensuring accuracy of the collected data
  - Verification of quality-of-service parameters
  - Data aggregation

2. Data analysis— Baselining and reporting

- Data analysis subtasks include
  - Network and device traffic characterization and analysis functions
  - Performance
  - Exceptions
  - Capacity analysis
  - Baselining
  - Traffic forecasting

3. Performance management—

- Whereas monitoring only observes activities performance management includes adjusting configurations to improve the network's performance and traffic handling (threshold definitions, capacity planning, and so on).

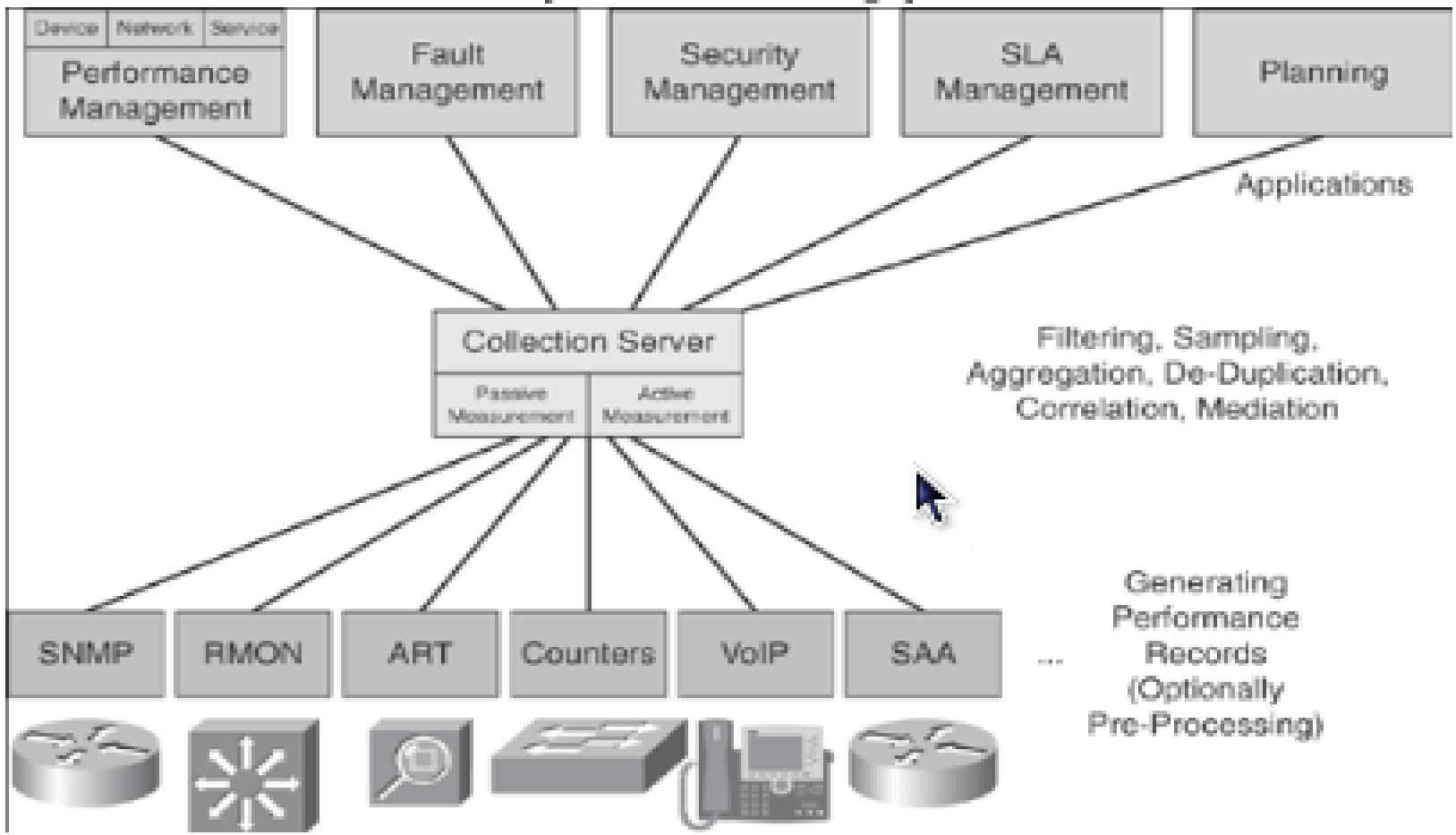
  - Management subtasks include
    - Ensuring compliance of SLAs and class-of-service (CoS) policies and guarantees
    - Defining thresholds
    - Sending notifications to higher-level applications
    - Adjusting configurations
    - Quality assurance

- Figure 1-4 shows the performance management architecture.,

- Performance management can also apply active measurements. In this case, we inject synthetic traffic into the network and monitor how the network treats it.

Figure 1-4. Performance Management Architecture

[View full size image]



A further refinement of performance management identifies three subcategories:

- Device-specific performance management
- Network-centric operations management
- Service management

- Device-specific performance management
  - considers the device in an isolated mode.
  - The device status is almost binary: it operates correctly, or a fault occurred.

- Performance monitoring at the network element level can also be considered binary, after the thresholds definition. For example,
  - if CPU utilization in the range of 5 to 80 percent is considered normal,
  - link utilization should be below 90 percent, and
  - interface errors should not exceed 1 percent.
  - Therefore, depending on whether the established threshold is exceeded, those situations are either normal or abnormal.

Network-centric performance management

- extends the focus to a network edge-to-edge perspective.

- Even though all devices might appear OK from a device perspective, the overall network performance might be affected by duplex mismatches, spanning-tree errors, routing loops, and so on.

- Service management
  - addresses the level above network connectivity.
- A service can be
  - relatively simple, such as DNS, or
  - complex, such as a transactional database system.
- the user expects the service to be
  - available and
  - have a predictable response time.

❖ needs to include service monitoring as well as management functions to modify components of the service in case of failures.

- performance takes into account details such as
  - network load, device load, throughput, link capacity, different traffic classes, dropped packets, congestion

- The collection interval.
- A data collection process for performance analysis should notify the administrator immediately if thresholds are exceeded; therefore, we need (almost) real-time collection in this case.

- Performance management needs history data to analyze deviation from normal as well as trending functions.

- Monitoring device health information, such as CPU, memory utilization, etc. is a crucial component of performance management

- Example 1.
- consider an otherwise normal network situation with average traffic load and then a user decides to install "interesting" software without notifying the administrator.
- E.g., installs a monitoring tool and starts discovering devices in the network.
  - SNMP communities may have been left set to the default values "public" and "private,". at the same time security restrictions (such as access control lists [ACL]) may not be in place,
- the user discovers network- and device-related details.
- The situation becomes critical when the user's monitoring tool collects the routing table of an Internet edge router.
  - For example, retrieving the complete routing table of a Cisco 2600 router with 64 MB of RAM and 4000 routes takes about 25 minutes and utilizes about 30 percent of the CPU.
- A performance-monitoring application would identify this situation immediately and report it.

Example 2 - a mis-configured link

- consider that a logical connection between two routers was configured as a trunk of three parallel links.
- For troubleshooting, the administrator shut down two of the links and then solved the issue.
- However, he put only one link back to operational, providing only two-thirds of the required bandwidth.
- Traffic would still go through

❖ but the increased utilization of the two active links should be identified by performance monitoring.

Example Actions

- In the first example,
  - performance management application could send a notification to a fault application and configure an ACL at the device to stop the unauthorized SNMP information gathering.
- In the second example,
  - performance management application could automatically activate the third link and notify the administrator.

- performance monitoring can be passive or active.

- Passive monitoring
  - gathers performance data by implementing meters.
  - Examples range from
    - simple interface counters to
    - dedicated appliances such as a Remote Monitoring (RMON) probe.

- Passive measurement needs to monitor some or all packets that are destined for a device.

  - It is called sampling if only a subset of packets is inspected versus a full collection if all packets are inspected.

- In some scenarios, such as measuring response time for bidirectional communications, implementing passive measurement can become complex, because
  - the request and response packets need to be correlated.

- An example is the Application Response Time (ART) Management Information Base (MIB), which extends the RMON 2 standard.
- ART measures delays between request/response sequences in application flows, such as HTTP and FTP, but it can monitor only applications that use well-known TCP ports.
  - To provide end-to-end measurement, an ART probe is needed at both the client and the server end. Cisco implements the ART MIB at the Network Analysis Module (NAM).

- The advantage of passive monitoring is that it does not interfere with the traffic in the network, so
  - ➢ the measurement does not bias the results.

- This benefit can also be a limitation, because network activity is the prerequisite for passive measurement.
  - Eg., observed traffic can indicate that a phone is operational, but how do you distinguish between an operational phone that is not in use and a faulty one if neither one generates any traffic?

  - ❖ Better to send some test traffic to the phone and monitor the results or alternatively have the device send keepalives regularly.

- The active monitoring approach
  - injects synthetic traffic into the network to measure performance metrics such as
    - availability, response time, network round-trip time, latency, jitter, reordering, packet loss, etc.

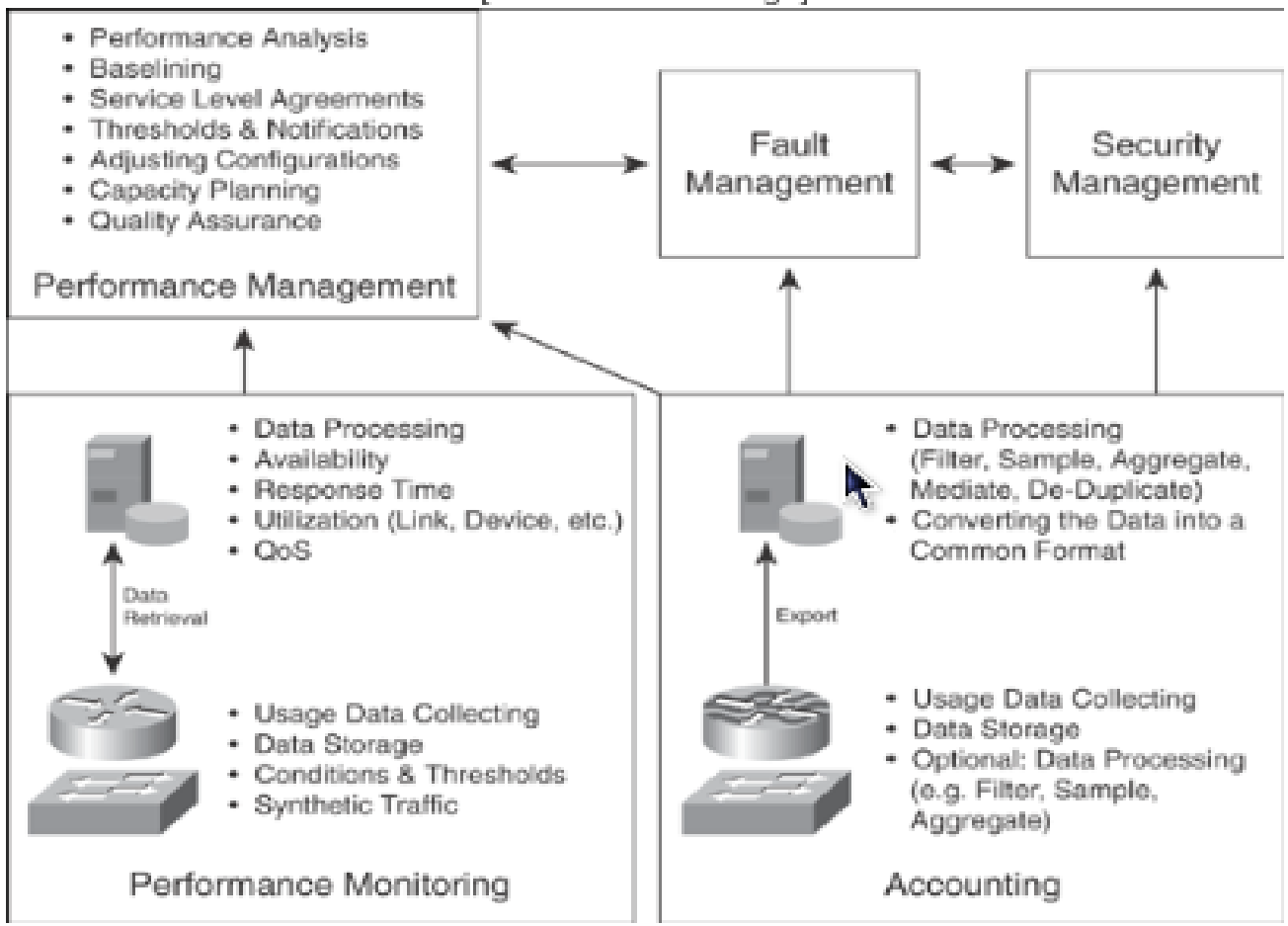
- The simplicity of active measurement increases scalability because only the generated traffic needs to be analyzed. The Cisco IP SLA is an example of active monitoring.

❖ Best current practices suggest combining active and passive measurements because they complement each other.

Best current practices suggest combining active and passive measurements because they complement each other.

Figure 1-5. Network Management Building Blocks

[View full size image]



Network Monitoring

- The term "network monitoring" is widely interpreted: one person might relate it to device utilization only, and someone else might think of end-to-end monitoring.
- In fact, network monitoring is a vague expression that includes multiple functions.
- Network monitoring applications enable a system administrator to monitor a network for the purposes of security, billing, and analysis (both live and offline).

- Table 1-2 illustrates device utilization.
- Assume that we have a network with three service classes deployed. Class 0 delivers real-time traffic, such as voice over IP, and
- class 1 carries business-critical traffic, such as e-mail and financial transactions.
- Class 2 covers everything else; this is the "best-effort" traffic class.
- Table 1-2 illustrates the total amount of traffic collected per class, including the number of packets and number of bytes.
- This report provides relevant information to a network planner.
- The technology applied in this example is an SNMP data collection of the CISCO-CLASS-BASED-QOS-MIB (see Chapter 4, "SNMP and MIBs"), which describes all the CoS counters.

Table

Table 1-2. Example of a Daily Report with Three Servicee Classes

	Class 0		Class 1		Class 2	
Time (Hour)	Packets	Bytes	Packets	Bytes	Packets	Bytes
0	38	2735	1300	59800	3	1002
1	55	3676	400	44700	61	9791
2	41	36661	400	16800	4	240
3	13	1660	200	8400	4	424
4	16	14456	400	44700	4	420
5	19	2721	400	44400	1	48
6	21	24725	600	35600	516	20648
7	19	3064	700	412200	15	677
8	5	925	1200	176000	1	48
9	4	457	1300	104100	1242	1489205
10	5	3004	1900	1091900	1	48
11	4	451	400	39800	545	22641
12	4	456	800	54200	1017	1089699
13	5	510	500	41600	36	3240
14	4	455	400	99300	15	3287
15	5	511	800	36800	685	27578
16	4	454	100	4000	3	144
17	4	457	500	309500	2	322
18	4	455	400	34100	4	192
19	5	3095	1300	104100	4	424
20	4	398	100	15200	4	424
21	5	1126	800	54200	12	936
22	7	782	1300	104100	4	835
23	9	7701	600	35600	1	235

Another scenario of network monitoring is the use of accounting usage resource records for performance monitoring. The accounting collection process at the device level gathers usage records of network resources. These records consist of information such as interface utilization, traffic details per application and user (for example, percentage of web traffic), real-time traffic, and network management traffic. They may include details such as the originator and recipient of a communication. Granularity differs according to the requirements. A service provider might collect individual user details for premium customers, whereas an enterprise might be interested in only a summary per department. This section's focus is on usage resource records, not on overall device details, such as CPU utilization and available memory.

- A network monitoring solution can provide the following details for performance monitoring:

- Device performance monitoring:
  - Interface and subinterface utilization
  - Per class of service utilization
  - Traffic per application
- Network performance monitoring:
  - Communication patterns in the network
  - Path utilization between devices in the network
- Service performance monitoring:
  - Traffic per server
  - Traffic per service
  - Traffic per application

- Applied technologies for performance monitoring include
  - SNMP MIBs,
  - RMON,
  - Cisco IP SLA, and
  - Cisco NetFlow services

Application Monitoring and Profiling

- A collection of application-specific details is also very useful for network baselining.
- Running an audit for the first time sometimes leads to surprises, because more applications are active on the network than the administrator expected.
- Application monitoring is also a prerequisite for QoS deployment in the network.
- To classify applications in different classes, their specific requirements should be studied in advance, as well as the communication patterns and a traffic matrix per application.
  - Real-time applications such as voice and video require tight SLA parameters, whereas e-mail and backup traffic would accept best-effort support without a serious impact.

- In most environments, applications fall into the following distinct categories:
  - Applications that can be identified by TCP or UDP port number. Either "well-known" (0 through 1023) or registered port numbers (1024 through 49151).
  - Applications that use dynamic and/or private application port numbers (49152 through 65535), which are negotiated before connection establishment and sometimes are changed dynamically during the session.
  - Applications that are identified via the type of service (ToS) bit. Examples such as voice and videoconferencing (IPVC) can be identified via the TOS value.

- The following list of applications and protocols comprises about 80 percent of the total traffic that traverses the WAN:
  - HTTP
  - E-mail
  - IP telephony
  - IP video
  - Server and PC backups
  - Video on demand (VoD)
  - Multicast
  - SNMP
  - Antivirus updates
  - Peer-to-peer traffic

- Techniques to obtain the classification per application are
  - RMON2,
  - Cisco NetFlow, and
  - Cisco NBAR.
- All three classify the observed traffic per application type.

- Best practice suggests monitoring the network before implementing new applications.
- Taking a proactive approach means that you analyze the network in advance to identify how it deals with new applications and whether it can handle the additional traffic appropriately.
- A good example is the IP telephony (IPT) deployment. You can run jitter probe operations with Cisco IP SLA, identify where the network needs modifications or upgrades, and start the IPT deployment after all tests indicate that the network is running well.
- After the deployment, accounting records deliver ongoing details about the newly deployed service.
- These can be used for general monitoring of the service as well as troubleshooting and SLA examination.

Capacity Planning

- Network traffic increases on a daily basis
- Different studies produce different estimates of how long it takes traffic to double.
- This helps us predict that today's network designs will not be able to carry the traffic five years from now.

- Bandwidth consumption may double every 18 months.

- This requires foresight and accurate planning of the network and future extensions.
- Enterprises and service providers should carefully plan how to extend the network in an economical way.

- A service provider might consider the following:
  - Which point of presence (PoP) generates the most revenue?
  - Which access points are not profitable and should be consolidated?
  - Should there be spare capacity for premium users?
  - In which segment is the traffic decreasing? Did we lose customers to the competition? What might be the reason?

- An enterprise IT department might consider the following:
  - Which departments are growing the fastest? Which links will require an upgrade soon?
  - For which department is network connectivity business-critical and therefore should have a high-availability design?

- These questions cannot be answered without an accurate traffic analysis; it requires a network baseline and continuously collected trend reports.

- Capacity planning can be considered from
  - the link point of view or from
  - the network-wide point of view.

  Each view requires a different set of collection parameters and mechanisms.

Link Capacity Planning

- For link capacity planning, the interface counters stored in the MIB are polled via SNMP, and the link utilization can be deduced.
- This simple rule of thumb is sometimes applied to capacity planning.
  - If the average link utilization during business hours is above 50 percent, it is time to upgrade the link!

to calculate utilization:

input utilization =

[(Δ(ifInOctets)) * 8 * 100] / [(number of seconds in Δ) * ifSpeed]

output utilization =

[Δ(ifOutOctets)) * 8 * 100] / [(number of seconds in Δ) * ifSpeed]

- Link Capacity Planning

- Some alarms, such as a trap or a syslog message, may be sent to the fault management application to detect a threshold violation.