Lecture 13. Server monitoring

Definition

Server monitoring is the process of gaining visibility into the activity on your servers — whether physical or virtual. Servers are devices (or increasingly, applications) that store and process information that is provided to other devices, applications or users on-demand. A single server can support hundreds or even thousands of requests simultaneously. As such, ensuring that all of an organization’s servers are operating according to expectations is a critical part of managing your IT infrastructure.

Server monitoring term

The term “server monitoring” is complex because of the exceptionally wide range of servers that exist. A web server can be a physical device, but it increasingly refers to a virtual server housed on a physical machine shared by dozens of other clients, each running their own independent web server system. Mail servers, print servers and database servers are just a few types of server devices and software.

Monitoring and alerting to issues on these various servers each requires a specific type of technological oversight, and the typical “off the shelf” server monitoring tool is unlikely to be appropriate for every one of them. In this article, we’ll help explain how various server monitoring tools and monitoring services work, the value they bring to the enterprise, and how to go about selecting the right system for your organization.

Importance of server monitoring

Servers are some of the most critical pieces of your IT infrastructure, so it stands to reason that monitoring their performance and uptime is vital to the health of your IT environment. If a web server is offline, running slowly, experiencing outages or other performance issues, you may lose customers who decide to visit elsewhere. If an internal file server is generating errors, key business data such as accounting files or customer records could be corrupted.

The exact workflow for server monitoring will change based on your chosen server monitoring software solution and the cloud-based server that you are trying to monitor. As your IT organization grows in size and number of deployments, you will need to select and configure a server monitoring tool that regularly collects data from every one of your cloud-based servers.

Server monitoring steps

The general process of server monitoring can be described in five steps:

1. **Identify the Most Important KPIs** - Server monitoring begins with the identification of exactly what data you would like to track on each server. Your choices here depend on the functionality that the server is delivering for your organization. For an application server, you might decide that the critical KPIs are availability and responsiveness. For a web server, capacity and speed might be the most important. For a data storage server, you might be more concerned about latency, data throughput, and data loss.
2. **Set Baseline KPI Values** - Once you have determined which KPIs are the most important, the next step is to measure the performance of each server on each KPI metric and determine an acceptable range of values for the KPI. This initial measurement will act as a baseline against which the future performance of the server will be measured.

3. **Configure Data Collection and Analysis** - A server monitoring tool must be appropriately configured to pull data from the servers deployed in your cloud environment. Server monitoring tools track the activity on the server by streaming event logs, also called log files, that the server automatically generates. Log files contain information about errors, user activity and security events that happen on the server. In addition to log files, server monitoring tools track server operating system KPIs including CPU and memory availability, network connectivity and disk performance.

4. **Set up Comprehensive and Specific Alerts** - Now that you have configured your data collection and aggregation, the next step is to build out an alert system that will send notifications to you and your team when there is a KPI breach and your chosen metrics drop below threshold levels.

5. **Get Ready to Respond** - Finally, you'll need to outline policy and procedure for responding to alerts. Who is responsible for investigating security alerts? Finding solutions to operational issues? What kinds of alerts should warrant a response, and how urgent should the response be? These are all questions that need to be answered as you define how your organization will treat each type of event notification.

Server type

Today's leading server monitoring software tools can pull event logs for many types of servers, including:

Web Servers - A web server is configured to deliver web pages. Web servers have a unique IP address and domain name that corresponds to the website they host.

Application Servers - IT organizations can identify operational efficiencies by tracking the health, performance, and load of applications that are deployed in the cloud. Commonly prioritized metrics for application server monitoring include resource usage, data throughput, the latency of responses, service failures and restarts, error rates and success rates and overall application availability.

Network Servers - A network server acts as a central hub, helping other machines in your network access additional computing resources like processing power, disk space or printers on an on-demand basis. Network servers can also be used to store files or run applications from a central location.

What is server performance monitoring?

While server monitoring is a broad term that concerns the overall health of a server, server performance monitoring is concerned strictly with performance metrics. For a physical server, metrics primarily include memory and CPU utilization, as well as disk I/O and network performance. For a virtual server, performance metrics may include database or web server response time, network bandwidth utilization, and other measures of resource utilization, depending on the specific type of server.

Service performance monitoring is important for a variety of reasons. First, it is often predictive in nature — slowdowns and other performance issues can be instructive in helping IT pinpoint problems that are developing. Bottlenecks can be useful in showing where component or service upgrades are needed, and capacity management tools can be used to project what resources may be needed to support a new application or other workloads.
Compliance is another big issue that informs server performance monitoring. Many enterprises are committed to providing a certain level of uptime or performance, which can be critical in high-stress environments such as financial trading, SaaS offerings, and streaming media. If performance falls below certain thresholds, compliance penalties can be severe.

What are best practices for server monitoring?

While every environment is different, key best practices can help to ensure your IT department gets the most out of their investment in a server monitoring solution.

- **Ensure hardware is operating according to appropriate tolerance levels:** File servers are often pushed to their operational limits, and very few ever get a break, running 24/7 with no room for any downtime. Pay careful attention to key metrics like CPU temperature, CPU and RAM utilization, and storage capacity utilization to ensure every server is always running at peak physical performance. These checks, called “heartbeat” checks, should be configured at regular intervals.
- **Proactively monitor software for failures:** Use your server monitoring tools to watch for software problems as well as hardware issues. For example, server monitoring tools can help alert you to errors that arise if a database has become corrupted, if a security event has disabled key services, or if a backup has failed.

- **Consider your history:** Server problems rarely emerge in a vacuum. Consider the historical context of any problems that arise by charting metrics over time — generally 30 days or 90 days. For example, has CPU temperature abruptly risen in the last few days? This could indicate a server fan is failing.
- **Keep tabs on alerts:** Alerts should be monitored in real time as they arise, then triaged and assigned to an analyst for a resolution. This is the most common way in which an analyst can determine that something has gone wrong. Find a reliable way to manage and prioritize the most critical alerts through the noise. When incidents are escalated, make sure it gets to the right person at the right time to ensure better team collaboration.

- **Use server monitor data to plan short-term cloud capacity:** In a virtual server scenario, your server monitoring system can be instrumental in helping to plan how much computing power you need at any given moment. If services begin to slow down for users or experience other performance issues, IT management can use the server monitor to assess the situation and quickly spin up additional resources — or take them offline, if demand is low.
- **Get a jump on capacity planning:** Datacenter workloads have roughly doubled over the last five years, and servers have had to keep up. By monitoring long-term trends in server utilization, you can be better prepared for future server needs (both online and off).
- **Expand asset management and tracking:** Server monitoring can give you insight into when systems are approaching end of life — or tell you if assets have vanished from the network altogether (often indicating either failure or theft). Instead of relying on spreadsheets to track physical hardware in the enterprise, let your server monitoring tool do the work for you.

Server monitoring capabilities

When considering a server monitoring tool, you’ll want to assess these key server monitoring capabilities:

- **Breadth of coverage:** Does the tool support all the server types (hardware and software; on-premises and cloud) that your enterprise uses? Is it prepared for future types of servers your enterprise may implement down the road?
- **Intelligent alert management:** Is it easy to set up alerts via the configuration of thresholds that trigger them? How are alerts delivered? Are mobile users a consideration?
- **Root cause investigation intelligence:** Does the tool include logic or AI algorithms to help you determine why a problem has occurred, rather than telling you that something has gone wrong without context?
- **Ease of use:** Does the system include an intuitive dashboard that makes it easy to monitor events, perform triage, and react to problems quickly?
- **Support policy:** How easy is it to get in touch with technical support if you need help?

The following outline is a list of items to take into account when implementing a server monitoring system:

What should you monitor?

- Monitor standardized operating system specific KPIs and use appropriate thresholds
- Monitor Operating System availability with pings
- Monitor the availability of server specific functions
- Monitor Event Logs on Windows and syslogs on Unix/Linux and network devices
- Server specific known problems (e.g. service or process crashes)

Thank you for your attention!