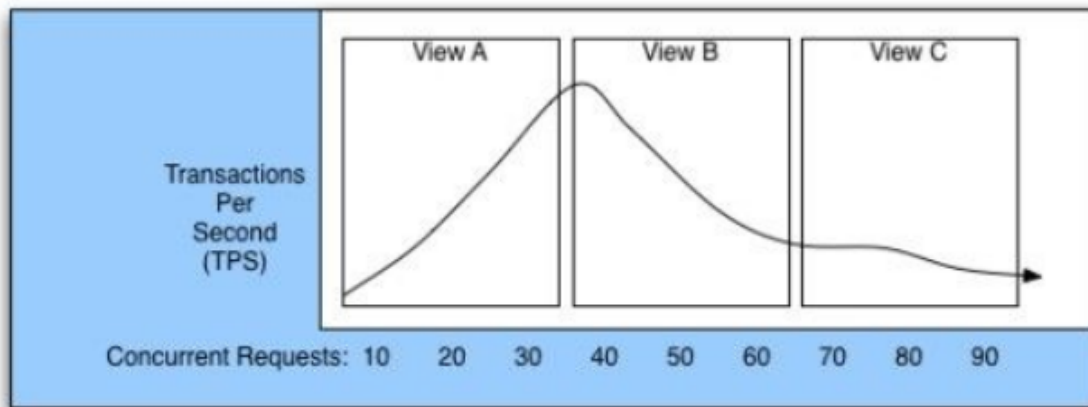


Scalability Index

The **Scalability Index** describes how a service responds to increasing concurrent request (CR) levels and payload sizes. It is used during calibration to identify the range in which the service operates efficiently and to determine appropriate load levels for the Full Test.

The chart below illustrates three representative regions of a scalability curve:



- **View A — Underdriving:** Increasing the CR level results in a corresponding increase in TPS. The service is still able to process additional load efficiently, indicating that the current load level may be below the service's optimal operating range.
- **View B — Optimal operating range:** The service reaches its most useful operating range. Increasing the CR level no longer produces proportional throughput gains, but the service continues to operate at an acceptable TPS level. This range is typically used to define the CR levels for the Full Test.
- **View C — Overdriving:** Increasing the CR level produces little or no additional TPS. The service may be saturated, or another system resource may have become the limiting factor. Increasing the load further is therefore unlikely to provide useful scalability information.

The goal of calibration is to identify the **View B range** rather than simply maximize the CR level. Testing at too few requests may underutilize the service, while testing at excessive levels may introduce saturation that obscures the service's useful scalability characteristics.

Once the appropriate CR range has been identified, use those levels in the **Full Test**. The same calibrated levels can also be used when comparing alternative APIs, implementations, or products so that their performance results can be evaluated under equivalent workloads.

Identifying Bottlenecks

Resource utilization can help determine what limits scalability during calibration. Depending on the type of service and test, review metrics such as:

- CPU utilization
- Memory utilization
- Network bandwidth
- Load-generator or consumer utilization
- Service response time and TPS

The relevance of each metric depends on the architecture and behavior of the service. For example, CPU and memory utilization can provide useful information for stateless services, while resource utilization may be less representative of scalability for stateful services where other factors determine the system's behavior.

The **Scalability Index should therefore be considered together with system-resource metrics and response-time behavior**, rather than relying on TPS alone.

Last updated: August 13, 2026

If you have any feedback on this topic, please contact us at documentation@appvance.ai.

Revision #5

Created 2026-07-31 15:09:10 UTC by Admin

Updated 2026-08-14 02:04:40 UTC by mmora