Diss. ETH No. 15233

Evaluating Performance in Systems with Heavy-Tailed Input
A Quantile-based Approach

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of
DOCTOR OF TECHNICAL SCIENCES

presented by
ULRICH FIEDLER
Dipl. Physiker ETH
born January 12, 1965
citizen of Germany

accepted on the recommendation of
Prof. Dr. B. Plattner, examiner
Dr. M. Maechler, co-examiner
Prof. Dr. P. Huang, co-examiner

2003

Abstract

One of the key invariants in computer and communication systems is that important characteristics follow long- or heavy-tailed distributions. This means that the tail of these distributions declines according to a power law. Hence, the probability for extremely large values is non-negligible. For example, such distributions have been found to describe the size of web objects or the processing latencies in computer and communication systems. As a consequence, there is a need to employ such distributions when evaluating such systems with synthetic workloads. However, sampling from such distributions to generate workloads implies that the system under evaluation remains in transient state over all periods of time that are feasible for performance evaluations. Consequently, frequently-used statistics for performance evaluation, such as the average of the system output, do not converge.

In this thesis we move away from evaluation using statistics such as the average, which describe the expected behavior of the system in all cases, and take the step towards evaluation using statistics such as quantiles, which describe the behavior in a given percentage of cases. Such quantiles of the system output do not depend on the extreme tail of the output distribution. We therefore address the problem of whether employing quantiles can enable performance evaluations within periods of time that are feasible in practice for performance evaluations.

Quantiles have a natural interpretation to statistically characterize the system performance. If e.g. a system offers a web service, the 99-th percentile of the latencies of downloads can statistically characterize the system performance from a user's view, since 99% of downloads terminate within times smaller than this quantile. If converged, such quantiles can be used to derive statistical guarantees for the system performance. Similar statements hold for

system components such as servers and networks.

Applying probability theory, we show that statistics of quantiles converge considerably faster than other frequently-used performance evaluation statistics if the underlying distribution is long- or heavy-tailed. Based on this theory, we give a method which enables to evaluate system performance under long- or heavy-tailed input within periods of time that are practically feasible. We validate the proposed method by applying it to a simulation-based evaluation of the network performance of systems that offer web services.

We show that the proposed method has further applications to other problems that require performance evaluation with synthetic workloads which are generated by sampling long- or heavy-tailed distributions. These applications include capacity provisioning, benchmarking of new hard- and software, as well the evaluation of protocols that rely on the request/reply paradigm such as HTTP, IMAP, FTP, or NFS. Further applications can be found in the field of computer systems where CPU requirements of tasks show a heavy-tailed distribution. This includes the evaluation of migration policies in a network of workstations, as well as task assignment policies for distributed servers.

Kurzfassung

Die Langschwänzigkeit der Verteilungen wichtiger Kenngrössen ist eine der bekannten Invarianten in Computer- und Kommunikationssystemen. Langschwänzigkeit einer Verteilung bedeutet, dass die Verteilung nach einem Potenzgesetz zerfällt. Dies wiederum hat zur Folge, dass die Wahrscheinlichkeit sehr grosse Werte zu beobachten nicht vernachlässigbar ist. Unter anderem wurde festgestellt, dass z.B. die Grösse von Web-Objekten oder die Verarbeitungslatenz gewisser Systeme mit langschwänzigen Verteilungen beschrieben werden können. Daher gibt es bei der Evaluation von Systemen mit synthetischer Last ein Bedürfnis solche Verteilungen zu modellieren. Allerdings impliziert das Generieren von Last durch ein Abtasten solcher Verteilungen, dass das zu evaluierende System sich in einem transienten Zustand befindet während aller Zeitspannen, die für Performance-Evaluationen machbar sind. Als Folge konvertieren gebräuchliche Statistiken zur Performancemessung, wie der Mittelwert von Outputgrössen, nicht.

In dieser Dissertation bewegen wir uns weg von Performance-Evaluationen mit Statistiken, wie dem Mittelwert, der das erwartete Verhalten des Systems in allen Fällen wiedergibt, und betrachten Statistiken wie Quantile, die das Verhalten des Systems in einem gegebenen Prozentsatz von Fällen wiedergeben. Solche Quantile der Outputgrössen hängen nicht vom äussersten Schwanz der Output-Verteilung ab. Deshalb untersuchen wir das Problem, ob das Betrachten von Quantilen eine Performance-Evaluation innerhalb von Zeiten ermöglicht, die für solche Evaluationen praktisch machbar sind.

Quantile haben eine natürliche Interpretation um die Performance von Systemen zu charakterisieren. Wenn z.B. ein System einen Web-Dienst anbietet, kann das 99-te Perzentil der Download-Latenz die System Performance aus

Benutzersicht charakterisieren, da 99% der Downloads innerhalb einer Zeit enden, die kleiner als dieses Quantil ist. Falls ein solches Quantil konvergiert hat, kann es benutzt werden, um statistische Garantien für die Performance des Systems abzugeben. Aehnliche Aussagen können für Systemkomponenten, wie Server und Netzwerke, gemacht werden.

Durch Anwendung von Wahrscheinlichkeitstheorie zeigen wir, dass Statistiken wie Quantile signifikant schneller als andere bei der Performance-Evaluation gebräuchliche Statistiken konvergieren, wenn die unterliegende Verteilung langschwänzig ist. Basierend auf dieser Theorie geben wir eine Methode an, die es ermöglicht, die System Performance unter langschwänzigem Input innerhalb von Zeiten zu evaluieren, die praktisch machbar sind. Wir validieren die vorgeschlagene Methode, indem wir sie auf eine simulationsbasierte Performance-Evaluation von Netzwerken in Systemen, die Web-Dienste anbieten, anwenden.

Wir zeigen, dass diese Evaluationsmethode weitere Anwendungen bei Problemen hat, die eine Performance-Evaluation mittels synthetischer Last erfordern, welche durch Abtasten von langschwänzigen Verteilungen generiert wird. Die Anwendungen beinhalten die Dimensionierung von Kapazitäten, das Benchmarking neuer Hard- und Software, sowie die Evaluation von Protokollen die auf dem Request/Reply Paradigma beruhen, wie HTTP, IMAP, FTP, oder NFS. Weitere Anwendungen können im Gebiet der Computersysteme gefunden werden, da der CPU-Bedarf von Tasks ebenfalls langschwänzig verteilt ist. Dies beinhaltet die Evaluation von 'Migration Policies' in einem Netzwerk von Workstations sowie von 'Task Assigment Policies' für verteilte Server.