

Federated Capacity Planning for Distributed Computing Infrastructures

Rechenleistung ist für beinahe alle Anwendungsgebiete in Wissenschaft und Industrie zum Gebrauchsgut geworden, und die Möglichkeit der Durchführung komplexer Berechnungen ist für wichtige Bereiche wie Biotechnologie, Klimaforschung und Produktdesign unverzichtbare Voraussetzung. Um diesem Bedarf zu genügen, sind in den letzten zehn Jahren unterschiedliche Konzepte zur flächendeckenden, verteilten Bereitstellung von Rechenleistung entwickelt worden, zu deren populärsten Vertretern wohl das akademische Grid Computing und das kommerzielle Cloud Computing gehören. Beiden Ansätzen ist dabei gemein, dass dem offensichtlichen Wunsch der Verbraucher – Forscher, Ingenieure – nach bedarfsgerechter und verlässlicher Nutzbarkeit von Rechenkapazität die Notwendigkeit einer einfachen und kostengünstigen Bereitstellung auf Seiten der Betreiber – akademische Hochleistungsrechenzentren, kommerzielle Data Center – entgegen steht.

In der vorliegenden Arbeit wird dieses Spannungsfeld im Bezug auf Kapazitätsplanung näher betrachtet. Im ersten Schritt wird hierfür ein geeignetes Modell entwickelt, welches die Erfordernisse moderner e-Infrastrukturen geeignet abbildet und die Erkenntnisse der letzten drei Jahrzehnte Forschung im Bereich des verteilten Rechnens einbezieht. Im Rahmen dieses Modells wird anschließend die Frage untersucht, ob – und falls ja, in welchem Umfang – die Bildung von föderierten Infrastrukturen aus Sicht der beteiligten Akteure (Nutzer und Anbieter) sinnvoll möglich ist, ohne die Einzelinteressen zu vernachlässigen. Hierbei werden Scheduling-Verfahren für unterschiedliche Architekturen entwickelt und bewertet, und die Grenzen des Erreichbaren ausgelotet. Schließlich wird evaluiert, welche Voraussetzungen auf Ebene der technischen Architektur erfüllt sein müssen, um die vorher untersuchten Algorithmen in echte Szenarien übertragen zu können. Zu diesem Zweck werden die Anforderungen zweier Produktionsumgebungen erhoben, und es wird ein allgemeines Interaktionsmodell entwickelt, welches den Einsatz der entwickelten Planungsverfahren erlaubt. Dabei wird mithilfe zweier neu entworfener Protokolle nicht nur die Übertragbarkeit der Methoden in reale Systeme gezeigt, sondern auch dargestellt, dass aufgrund der Erweiterbarkeit der Protokolle und Architektur eine Vielzahl von weitergehenden Anwendungsszenarien umsetzbar ist.

Schlagnworte: Scheduling, Resource Management, Distributed Computing, Grid Computing, Cloud Computing, Architektur, Standards