Power Delivery in Data Centers

Milan Stephan
Ausgewählte Kapitel der Systemsoftware (WS 2020/21)
Friedrich-Alexander-Universität Erlangen-Nürnberg
milan.stephan@fau.de

ZUSAMMENFASSUNG

Um das Leistungsbudget eines Rechenzentrums besser ausnutzen zu können, ist eine möglichst genaue Analyse des Verbrauchs der verschiedenen Geräte sinnvoll. Darauf aufbauend können Einsparungen bei vorhandenen Systemen vorgenommen und die verfügbar gemachte Leistung anderweitig verteilt werden. Diese Ausarbeitung thematisiert die Variationen des Energieverbrauchs, stellt Ansätze zur Reduzierung und Umverteilung vor und bewertet diese.

1 EINLEITUNG

Die größten laufenden Kosten für Rechenzentren setzen sich aus der Energie für die Server und die Klimatisierung zusammen. Da Rechenzentren mit diesen Kosten haushalten müssen, ist eine Maximierung der Rechenkraft bei gegebener elektrischer Leistung erstrebenswert. Dazu haben Rechenzentren oft feste Leistungsbudgets, die nicht überschritten werden dürfen. Optimal wären so genannte „Energie-proportionale“ Systeme, deren Gesamtverbrauch proportional zu ihrer geleisteten Rechenarbeit ist [1]. Da es diese noch nicht gibt, kann man lediglich versuchen, die Hardware bei möglichst konstanter Last nahe ihres Effizienzmaximums zu betreiben.

Prozessoren skalieren dahingehend bereits einigermaßen gut und können Teile ihrer inneren Strukturen bei Nichtgebrauch ohne größere Nachteile abschalten. Andere Komponenten wie Festplatten haben auf der anderen Seite recht hohe Latenzen, um aus dem ausgeschalteten Zustand wieder betriebsbereit zu sein. Hier ist das Einsparpotenzial aufgrund dieser Nachteile weniger attraktiv.

Diese Ausarbeitung thematisiert die verschiedenen Verbraucher in Rechenzentren, stellt Optimierungsansätze für eine bessere Energieeffizienz vor und bewertet diese. Dabei liegt der Fokus auf hardwarenahen Optimierungen ohne akkurates Wissen über die betriebenen Applikationen vorauszusetzen. Zudem werden mit einem Ansatz die Vorteile von kooperierender Software aufzeigt, die der Energieverwaltung Metriken bereitstellt.

2 GRUNDLAGEN

Der Leistungsbedarf eines Rechenzentrums lässt sich grob in Server, Kühlung, Aufbereitung der elektrischen Energie, Netzwerk und Beleuchtung einteilen. In dieser Ausarbeitung konzentriere ich mich auf die ersten beiden Aspekte, die zusammen etwa 85% des Leistungsbedarfs ausmachen [2, 3].

Die Aufbereitung der elektrischen Energie umfasst Verluste durch Transformation und den Einsatz von Unterbrechungsfreien Stromversorgungen (USV). Diese Geräte können kurzzeitige Versorgungsengpässe, häufig mittels Akkumulatoren, überbrücken, bis langfristigere Lösungen wie etwa Dieselgeneratoren gestartet werden.

2.1 CPU

Bei der CPU zeigt sich eine vergleichsweise gute Skalierung. Für die minimale Frequenz von 800 MHz eines AMD Ryzen 1700X beträgt die Leistungsaufnahme pro Kern unter einem Watt (unter Last). Mit steigender Frequenz nimmt sie allerdings näherungsweise quadratisch zu bis zu einem Maximum von etwa 12 W. Die Sprünge in Abbildung 1 [4] lassen sich mit den nur drei verfügbaren P-States des Ryzen 1700X erklären. Wird eine bestimmte Frequenz überschritten, so muss eine höhere Spannung angelegt werden, die wiederum die Leistungsaufnahme gemäß $P \propto V^2 f$ beeinflusst. Da die Spannung aufgrund der nur drei möglichen Spannungslevel in dem Beispiel des Ryzen 1700X nicht auf jede Frequenz abgestimmt werden kann, fallen manche Übergänge wie der von 3,4 auf 3,5 GHz besonders stark ins Gewicht. Das sind die so genanngen Turbo Boost Frequenzen. Jeder P-State muss alle Frequenzen in seiner Gruppe unterstützen können, daher könnte man mit fein granularen P-States weitere Einsparungen vornehmen. Die geleistete Arbeit pro Taktzyklus ist bei

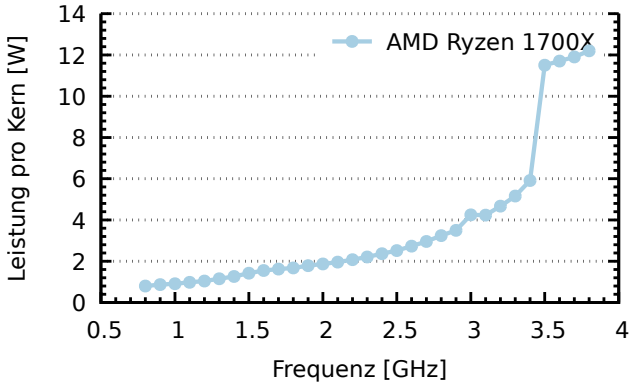

Abbildung 1: Leistungsaufnahme eines AMD Ryzen 1700X Kerns bei verschiedenen Betriebsfrequenzen unter Last [4]. Hoher Anstieg der benötigten Leistung bei Verwendung des Turbo Boosts.

CPU-lastigen Aufgaben wie dem hier gemessenen SPEC2017 nahezu konstant [4]. Dementsprechend skaliert die geleistete Rechenarbeit auch direkt mit der Frequenz des Prozessors.

Setzt man die Leistungsaufnahme mit der Menge der geleisteten Rechenarbeit ins Verhältnis, wie in Abbildung 2 dargestellt, so zeigt sich die höchste Effizienz im unteren Mittelfeld. Dieser Punkt gilt für genau diese CPU und wird

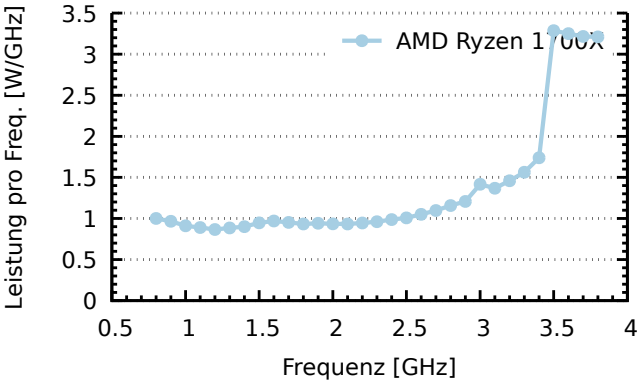

Abbildung 2: Effizienz eines AMD Ryzen 1700X Kerns bei verschiedenen Betriebsfrequenzen unter Last, basierend auf Abbildung 1. Im Turbo-Boost wird unverhältnismäßig viel Leistung benötigt.

hier stellvertretend angenommen. Selbstverständlich verhalten sich Modelle von Intel oder aus anderen Modellreihen ggf. stark abweichend. Das Augenmerk liegt hierbei auf der stark ansteigenden Leistungsaufnahme bei höheren Frequenzen. Aktuelle Prozessoren erreichen häufig Taktraten jenseits der 4 GHz, sodass diese Skalierung weiter an Bedeutung gewinnt.

Zur dynamischen Anpassung der CPU-Leistungsaufnahme an die zu leistende Arbeit und zur Einhaltung eines Leistungsbudgets stellt Intel „Dynamic Voltage and Frequency Scaling“ (DVFS) vor. Mit diesem Konzept kann sowohl die Leistung begrenzt werden, sowie die Spannung und Frequenz des Prozessors geregelt werden, um in vorzugsweise effizienten Regionen zu agieren. Turbo Boost von Intel bzw. eXtended Frequency Range (XFR) von AMD bieten darüber hinaus eine von Leistungsaufnahme und Temperatur abhängige automatische Übertaktung des Prozessors an. Diese kann bei kurzzeitigen Lastspitzen, wie sie beispielsweise auf Webservern vorkommen, die Latenzen verkürzen. Für den Dauerbetrieb sind diese Techniken allerdings aufgrund der überproportional hohen Leistungsaufnahme nicht zu empfehlen.

Um mehrere gleichartige Datensätze simultan zu verarbeiten, bieten moderne CPUs mit SSE und AVX Vektorinstruktionen an. Diese Instruktionen können bis zu 512 Bit, aufgeteilt auf 32- oder 64-Bit Werte, gleichzeitig berechnen. Allerdings benötigt die Ausführung dieser Instruktionen gegenüber einer Skalarrechnung mit nur einem Wert deutlich mehr Energie, weshalb die Frequenz bei der Ausführung solcher Instruktionen gesenkt wird. Dennoch bieten Vektorinstruktionen drastische Vorteile in der Verarbeitungsgeschwindigkeit geeigneter Datensätze [5].

2.2 Hauptspeicher

Große Mengen an Hauptspeicher können ebenfalls signifikanten Einfluss auf den Energieverbrauch eines Servers haben. Klassischer DDR-SDRAM benötigt regelmäßiges Auffrischen der Speicherzellen, um die Daten verfügbar zu halten [6].

So kann der Speicher in der Größenordnung von wenigen TiB mehr Leistung benötigen als die Prozessoren eines Servers [7]. Im Leerlauf eines Servers kann man die Speicherkommunikation abschalten und den Speicher im „Self-Refresh“ Modus laufen lassen, was laut Meisner et al. bis zu 90% Leistung einspart [8]. Diese Messungen gehen allerdings auf DDR2-SDRAM zurück, sind also keineswegs aktuell. Ebenso kommen neuere Speichertechnologien wie DDR4-SDRAM mit deutlich größeren Packungsdichten und geringeren Spannungen (1,2V statt 1,8V) einher, was eine geringere Leistungsaufnahme pro GiB bedeutet. Des weiteren gibt es mittlerweile auch nicht-volatile Speichertechniken, die komplett ohne Wiederauffrischung auskommen und den Energieverbrauch weiterhin drastisch senken können [9].

2.3 Festplatten / Solid-State-Laufwerke

Bei persistenten Speicherlösungen muss man zwischen klassischen Festplatten und Solid-State-Laufwerken unterscheiden. Klassische Festplatten speichern Daten auf rotierenden, magnetisierbaren Scheiben. Aufgrund der Rotation der Datenträger ergeben sich Latenzen beim Lesen und Schreiben von Daten. Mit größeren Rotationsgeschwindigkeiten können diese Latenzen zwar reduziert aber nicht ausgeschlossen werden. Zudem steigen damit die Reibungsverluste, die sowohl die Wärmeentwicklung, als auch die Leistungsaufnahme beeinflussen [10]. Ein Abschalten von Festplatten kommt aufgrund der hohen Latenzen bis zur erneuten Verfügbarkeit der Daten oft nicht in Frage. Das wird durch ausfallsicherere Systeme wie RAID [11], bei denen für den Zugriff auf ein Datum u.U. mehrere Festplatten herangezogen werden müssen, weiter verstärkt.

Mit Solid-State-Laufwerken hingegen können Daten mit viel geringeren Latenzen gelesen und geschrieben werden. Die auf Flash-Speicherzellen basierenden Datenträger können bei wahlfreien Zugriffen deutliche Geschwindigkeitsvorteile verbuchen. Zudem benötigen sie deutlichen weniger Energie im Leerlauf, und auch unter Last sind sie (pro Zahl an Anfragen) sparsamer als klassische Festplatten [12]. Allerdings sind SSDs pro GiB die deutlich kostspieligere Wahl [13].

2.4 Netzteile

Zur Energieversorgung der einzelnen Server werden Netzteile eingesetzt, die die Netzspannung auf die Betriebsspannung der Server transformieren. Jeder physische Server hat üblicherweise seine eigenen Netzteile, die vom Hardwarehersteller vorgesehen sind. Da Serversysteme in verschiedenen Ausstattungen angeboten werden, kann der Leistungshunger auch stark variieren. Beispielsweise lassen sich einige Server mit mehreren TiB an Hauptspeicher ausstatten und bieten ggf. auch mehrere Sockel für CPUs an. Dazu kommen optionale Beschleunigerkarten wie GPGPUs, die ebenfalls vom Netzteil mit versorgt werden müssen. Sollen alle diese Szenarien abgedeckt werden, so muss ein Netzteil im Stande sein, den maximal erwarteten Verbrauch der Hardware abdecken zu können. Häufig wird auch auf redundante Netzteile gesetzt, um Server trotz des Ausfalls eines Netzteils weiter betreiben

zu können. Bei Servern, die eher am unteren Ende des möglichen Verbrauchs angesiedelt sind, sind Netzteile daher kaum ausgelastet. Ein vorhandenes redundantes Netzteil verbraucht bei Nichtbenutzung ebenfalls Energie.

Netzteile erreichen üblicherweise eine Effizienz von 80% ab etwa 40% Nennleistung [8]. Unter diesem Schwellenwert lässt die Effizienz stark nach und es wäre sinnvoller, weniger leistungsstarke Netzteile zu verwenden.

2.5 Kühlung

Grundsätzlich lässt sich die Kühlung in Rechenzentren in zwei große Aspekte untergliedern. Die Kühlung der einzelnen Server deckt passive Kühlung über Kühlkörper und die aktive Kühlung mit Lüftern ab, die den Luftstrom über die Kühlrippen lenken, um die Wärme abzuführen. Kühlrippen stellen eine große Oberfläche bereit, um die Wärme an die umgebende Luft abzugeben. Durch den Luftstrom der Lüfter wird die erwärmte Luft abtransportiert. Dabei hat sowohl die Temperatur der einströmenden Luft, als auch der Luftdurchsatz der Lüfter einen großen Einfluss auf die abtransportierbare Wärmeenergie. Der Luftdurchsatz hängt von Anzahl und Art der verwendeten Lüfter ab, sowie von der Lüfterdrehzahl [14].

Um die erwärmte Luft wieder abzukühlen, wird eine Klimatisierung im Rechenzentrum verwendet. Typischerweise werden so genannte Computer Room Air Handler (CRAH) eingesetzt, in denen die warme Luft mit Hilfe von weiteren Lüftern durch wassergekühlte Spulen geleitet wird, bevor es erneut zur Kühlung der Server eingesetzt wird. Das durch die CRAHs erwärmte Wasser wird wiederum über eine Kältemaschine/Wärmepumpe gekühlt, die die Wärmeenergie in einen weiteren Wasserkreislauf und schließlich über einen Kühlturm in die Umwelt abgibt [14]. Selbstverständlich gibt es je nach den regionalen Gegebenheiten auch andere Möglichkeiten, eine effektive Kühlung zu realisieren.

Der Energieverbrauch der Lüfter wird somit von der Temperatur der einströmenden Luft, sowie der Zieltemperatur der Hardware bestimmt. Hier gilt es, einen Kompromiss zwischen dem Leistungsbudget der Lüfter in den Servern und dem Budget der Klimaanlage des Rechenzentrums auszuloten.

3 EINSPARUNGEN

Im Folgenden werden verschiedene Ansätze vorgestellt, die der Senkung der Leistungsaufnahme dienen. Dabei werden Möglichkeiten zur Reduktion der Anzahl der betriebenen Server betrachtet und Optimierungsmöglichkeiten pro Server selbst beleuchtet.

3.1 Konsolidierung

Der große Teil von Servern läuft im Durchschnitt bei einer Rechenlast von etwa 10% bis 50% [1]. Bei geringer Last benötigen Prozessoren zwar signifikant weniger Energie, der Grund-Leistungsbedarf eines kompletten Servers ist aber für das Umsetzen von nahezu keiner Rechenarbeit unverhältnismäßig hoch. Diesen Grundbedarf kann man durch Virtualisierung signifikant senken. Physische Server werden auf Virtuelle Maschinen (VMs) abgebildet, die dann auf Host-Systemen ausgeführt werden. Da sich eine Vielzahl von VMs ein Host-System teilen können, sinken die statischen Energiekosten pro Server mit steigender Anzahl an VMs auf dem Host. Bei einer Konsolidierung von 10 wenig ausgelasteten (etwa 6% CPU) VMs pro Host gehen die Autoren von einer Leistungseinsparung von etwa 87% aus [15].

Dennoch muss sichergestellt werden, dass die Host-Systeme entsprechend leistungsstark sind und dass die virtuellen Server genügend Ressourcen des Hosts bekommen können. Ein überlastetes Host-System würde die Leistungsfähigkeit aller darauf laufenden VMs reduzieren.

Für eine flexible Zuteilung von Ressourcen kann man sich die Möglichkeit der Migration von virtuellen Maschinen zu Nutze machen. Diese Technik erlaubt es den Betreibern des Rechenzentrums VMs auch im laufenden Betrieb auf andere Hosts umziehen zu können. Das ermöglicht neben den Leistungseinsparungen im Leerlauf auch die Minimierung von Ausfallzeiten, falls ein Host-System für Wartungsarbeiten heruntergefahren werden muss. Bei hoher Auslastung der VMs ließen sich auch neue Host-Systeme starten, auf die die VMs anschließend verteilt werden können, um eine gleichmäßige Auslastung aller Systeme zu gewährleisten [16].

Dazu können VMs automatisiert erstellt und gestartet werden, da keine Hardware zusammengebaut oder angeschlossen werden muss. So können bei Lastspitzen schnell neue Instanzen gestartet werden, was bei physischen Servern manuelle Eingriffe erfordern würde. Virtuelle Maschinen mit keinem festen Host-System sind prädestiniert für die Anbindung an ein ans Netzwerk angeschlossenes Speichersystem.

3.2 Netzwerkspeicher

Anstatt Festplatten in jedem physischen Server zu betreiben, kann man durch Netzwerkspeicher dedizierte Systeme zur Datenspeicherung über das Netzwerk bereitstellen. Das bietet den Vorteil, dass man anstatt kleinerer Festplatten pro Server auf größere Festplatten zurückgreifen kann, deren Gesamtkapazität von den Servern anteilig genutzt werden kann. Die zentrale Speicherung erlaubt es, tatsächlich auch die Menge an Speicher bereitzustellen, die benötigt wird.

Des weiteren ist RAID1 durch die mehrfache Speicherung von Daten nicht speicher-effizient [11] und kann in größeren Systemen von besser geeigneten Technologien abgelöst werden. Statt beispielsweise zweier 1TiB Festplatten im RAID1 Verbund kann so für diesen Server Speicher reserviert werden, dessen nutzbare Gesamtgröße mit einem TiB nur einen Bruchteil einer viel größeren Festplatte ausmacht. Dadurch kann die Zahl der eingesetzten Festplatten und damit auch der Energiebedarf massiv reduziert werden. Auch lässt sich die Ausfallsicherheit auf komplexeren Speichersystemen mit signifikant weniger Overhead realisieren [17].

Je nach Anwendungsfall können hierbei etablierte Verfahren wie RAID6 verwendet werden, während verteilte Anwendungen meistens ohnehin von verteilten Speicher-Lösungen profitieren [18]. Damit kann sowohl die Datensicherheit erhöht, als auch die Zahl der Festplatten aufgrund größerer

Modelle reduziert werden, um den gleichen Platz bereitstellen zu können [19]. Battles et al. erreichen eine Einsparung von gut 80% gegenüber direkt angebundenen Festplatten.

3.3 Spurt zum Schlafenlegen

Am sparsamsten ist Hardware im (teil-)ausgeschalteten Zustand. Moderne Chips bieten auch Möglichkeiten der Abschaltung einzelner Kerne an, allerdings ist die autmatische Abschaltung und Skalierung der Spannung und Frequenz nicht optimal.

Mit dem „Race to sleep“ Ansatz wird versucht, die Zeit im Sleep-State zu verlängern und eintreffende Arbeitspakete möglichst schnell abzuarbeiten. Im Gegensatz dazu steht das Dynamic Voltage and Frequency Scaling (DVFS), das den Prozessor die Arbeitspakete bei niedrigeren Frequenzen abarbeiten lässt, jedoch mit weniger Gelegenheiten zum Schlafenlegen [8]. Für Web-Anwendungen, die üblicherweise kurze Lastspitzen ausmachen, liegen die Einsparungen mit diesem Ansatz bei etwa 70%.

Zusätzlich werden andere Komponenten ebenfalls in Schlafzustände versetzt, aus denen ein schnelles Aufwachen möglich ist. Hier wird beispielsweise der Hauptspeicher mit erfasst, der seine Daten durch „self-refresh“ beibehält. Die Latenzen zum Aufwachen liegen bei dem Ansatz im Mikrosekundenbereich.

3.4 Bessere Netzteilauslastung

Im Weiteren beschreiben die Autoren noch eine Strategie um die Energieversorgung von Servern zu optimieren. Statt überdimensionierter Netzteile, die bei geringer Auslastung nicht mehr effizient arbeiten, kann eine Vielzahl von kleinen Netzteilen verwendet werden, die bei Bedarf zugeschaltet werden können. Bei dem Einsatz in Web-Anwendungen konnten die Autoren weitere Leistungseinsparungen -aufbauend auf den Optimierungen in Abschnitt 3.3- von 26% verbuchen [8].

3.5 Aufteilung

Stellt man die geleistete Rechenarbeit eines Servers dem Energiebedarf gegenüber, fällt zum einen der bereits erwähnte Grundbedarf auf. Darüber hinaus bieten moderne Prozessoren eine dynamische Anpassung der Frequenz und Spannung, um den Aufgaben besser gerecht werden können. Bei wenig Last kann die Frequenz reduziert werden, um Energie einzusparen. Um Aufgaben schneller erledigen zu können, kann die Frequenz allerdings auch angehoben werden. Der Leistungsbedarf eines Prozessors hängt nach $P \propto V^2 f$ sowohl von der Frequenz, als auch von der für diese Frequenz benötigen Kernspannung ab. Nach Abbildung 2 sind diese so genannten Boost- oder XFR-Frequenzen nicht energieeffizient, für schlecht parallelisierbare Aufgaben jedoch ggf. nötig.

Für skalierbare Anwendungen kann es daher Effizienzvorteile bringen, wenn anstatt eines Servers im Boost/XFR-Zustand weitere Kerne, ggf. auch auf mehreren Servern verwendet werden, die auf geringerer Frequenz arbeiten. Hierfür kann man statt der geleisteten Arbeit für CPU-lastige Arbeiten auch die Frequenz als Maßstab nehmen.

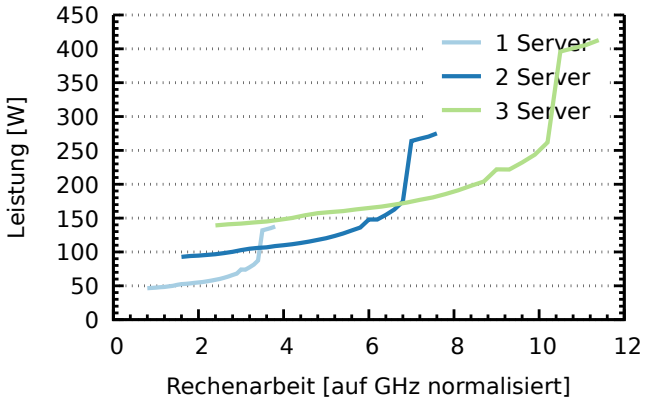

Abbildung 3: Rechenleistung (Anzahl gleicher Server · CPU-Frequenz jedes Servers) im Verhältnis zur elektrischen Leistung bei dem Einsatz von mehreren Servern

In dem in Abbildung 3 aufgeführten Beispiel wird ein Server mit dem Grundverbrauch von 40W angenommen. Bei voller Auslastung der CPU schneidet die Kurve des Szenarios mit einem Server den Fall mit zwei Servern. Ab dieser Frequenz benötigt der einzelne Server mehr Leistung als zwei Server, die die Rechenlast untereinander aufteilen. Statt den einzelnen Server auf maximaler Boost-Frequenz zu betreiben ist es für skalierbare Anwendungen daher effizienter, einen zweiten Server einzusetzen und die Frequenz bei beiden unterhalb der Boost-Zustände einzustellen.

3.6 Höhere Umgebungstemperatur

Um die Betriebstemperatur von Servern einzuhalten, werden zur aktiven Kühlung in den Servern Lüfter verwendet, die Wärme in die Umgebungsluft abführen. Dabei kann von der Temperaturdifferenz $T_{\text{Hardware}} - T_{\text{Luft}}$ jedoch nur ein Teil abgeführt werden. Je größer diese Temperaturdifferenz ist, desto mehr thermische Energie kann bei gleichem Lüftereinsatz abgeführt werden. Um diese Differenz zu erhöhen, kann man einerseits die Klimaanlage des Rechenzentrums heranziehen, allerdings benötigt diese zur stärkeren Kühlung ebenfalls mehr Leistung. Ein anderer Ansatz beschäftigt sich mit der Erhöhung der erlaubten Betriebstemperatur von Servern. So kann ohne die Erhöhung der Kühlungskosten die Menge der abgeführten Wärme verbessert werden. Hierbei ist allerdings zu beachten, dass durch die höhere Temperatur vermehrt Leckströme auftreten, die die Effizienz der Prozessoren senken. Ziel ist es, ein Optimum zu finden, bei dem die Leistung für die Kühlung bei gleicher Rechenarbeit minimiert werden kann [3].

4 UMVERTEILUNG

Mit der eingesparten Leistung, wie zuvor beschrieben, kann man nun weitere Server versorgen. Alternativ kann man auch besonders kritische Infrastrukturen mit höheren Leistungsbudgets ausstatten, um beispielsweise Zeitgarantien einhalten zu können.

4.1 Spannungs- und Frequenzskalierung

Die Spannungs- und Frequenzskalierung von A. Guliani und M. M. Swift stellt einen Ansatz vor, um die Leistungsaufnahme pro Anwendung dynamisch zu regulieren [4]. Zuerst gilt es, die Skalierung einer Anwendung mit der Leistungsaufnahme eines Prozessors zu ermitteln. Beispielsweise benötigt ein System zur persistenten Speicherung in der Regel weniger Prozessor-Ressourcen, als eine physikalische Simulation. Hier ist zu analysieren, wie viel Rechenleistung Anwendungen bei reduzierter Frequenz tatsächlich einbüßen. Ebenso gibt es Aufgaben, die je nach verfügbaren Ressourcen im Hintergrund ausgeführt werden können und solche, die unter allen Umständen Vorrang genießen sollen. Nach diesen Kriterien kann man Anwendungen in vier Gruppen einteilen:

- HDHP: hohe Anforderungen, hohe Priorität
- HDLP: hohe Anforderungen, niedrige Priorität
- LDHP: niedrige Anforderungen, hohe Priorität
- LDLP: niedrige Anforderungen, niedrige Priorität

Das Leistungsbudget eines Prozessors sorgt typischerweise bei hoher Rechenlast auf allen Kernen für eine niedrigere Frequenz im Vergleich zu einigen untätigen Kernen. Techniken wie Intels Turbo Boost und AMDs XFR machen sich diese Möglichkeiten zur Umverteilung zu Nutze. Im Umkehrschluss muss man bei einem Prozessor im Leistungslimit auf allen Kernen eine niedrigere Frequenz in Kauf nehmen oder einzelne Kerne selektiv drosseln, sodass die übrigen Kerne mehr Leistung zur Verfügung haben. Durch geschickte Einplanung von verschiedenen Typen von Anwendungen kann man hierbei die geleistete Rechenarbeit maximieren. So skaliert die erbrachte Arbeit einiger Anwendungen weniger mit den Prozessorressourcen als beispielsweise den Übertragungsgeschwindigkeiten ins Netzwerk oder zu Datenträgern. Wieder andere Applikationen profitieren zwar sehr stark von einer höheren Frequenz, lasten den Prozessor dabei aber ggf. weniger aus als andere.

Die drei Ansätze, nach denen man die Leistungseinteilung vornehmen kann, sind elektrische Leistung, Frequenz und Rechenleistung. Bei der elektrischen Leistung bekommen Anwendung einen Anteil des Leistungsbudgets vorgegeben, das sie ausschöpfen dürfen. Sind beispielsweise zwei Anwendungen eingelastet, so könnte eine zwei Drittel, die andere ein Drittel der Leistung ausnutzen. Da Anwendungen sehr unterschiedlich mit der Leistung skalieren können, ist dieser Ansatz noch ausbaufähig.

Mit der frequenzbasierten Einteilung wird nicht die benötigte Leistung, sondern die Prozessorfrequenzen für die einzelnen Anwendungen ins Verhältnis gesetzt. Auch hier ist die Skalierung maßgeblich für die Effektivität dieses Ansatzes.

Ideal ist eine Ressourcenaufteilung basierend auf der aktuell erreichten zu der maximal möglichen Rechenleistung einer Anwendung. So kann man auch unterschiedlich stark skalierende Anwendungen fair auf einem System laufen lassen. Beispielsweise lassen sich so auch bei eingeschränktem Leistungsbudget mehrere Anwendungen unterhalten, die bei 90% Rechenleistung arbeiten. Allerdings müssen hierfür Metriken über die Leistungsfähigkeit einer Anwendung erfasst werden können, die mit in die Leistungsverteilung einfließen. Für alle drei Ansätze müssen Anwendungen auf Kerne „gepinnt“ werden, damit der Scheduler des Betriebssystems keine ungeplanten Entscheidungen vorwegnimmt.

4.2 Energiepuffer

Ein weiterer Ansatz zur besseren Ausnutzung des Gesamt-Leistungsbudgets ist die Zwischenspeicherung von Energie. Die Leistungsaufnahme von Servern und damit auch des gesamten Rechenzentrums ist nicht konstant sondern unterliegt Schwankungen, sowohl über den Tag verteilt, als auch tagesabhängig. Somit werden Server, die beispielsweise Freizeit-Anwendungen betreuen, durchschnittlich gesehen mehr zu arbeitsfreien Zeiten wie dem Abend oder an Wochenenden belastet. Unternehmerische Server-Infrastruktur wird hingegen für eine höhere Auslastung zu Geschäftszeiten sorgen. Mittels Energiespeicherung kann man zu Zeiten, in denen weniger Leistung für Server benötigt wird, Energie vorhalten. Diese gepufferte Energie kann anschließend zu Zeiten eines höheren Verbrauchs abgerufen werden [20, 21].

Hier ist die richtige Dimensionierung essentiell für eine effiziente Energieversorgung, da auch bei der Speicherung Verluste auftreten und Speicherlösungen sowohl Platz, als auch finanzielle Mittel benötigen. Die Speicherung von Energie lässt sich vermutlich auch mit dem vorherigen Ansatz kombinieren, um die Leistungsbudgets der Server zusätzlich an die noch verbleibende Energie im Speicher zu koppeln.

5 EINSCHÄTZUNG

Die vorgestellten Ansätze können die Leistungsaufnahme verschiedener Komponenten eines Rechenzentrums drastisch senken, sind allerdings nicht alle sinnvoll oder mit geringem Aufwand zu implementieren.

Zur Minimierung von leer laufenden Servern trägt die **Konsolidierung** der Serverhardware zu VMs bei. Geeignete Systeme sind Server mit geringer Last, bei denen sich der Energieverbrauch hauptsächlich auf den konstanten Anteil beläuft. Dieser Ansatz kann sowohl die Energiekosten drastisch reduzieren, als auch die Menge und damit die Kosten der benötigen Serverhardware. Da sich die virtuellen Server Hardware teilen müssen, kann es jedoch Einbußen bei der Leistungsfähigkeit der einzelnen Server geben. Daher ist dieser Ansatz eher ungeeignet für Server, die dauerhaft sehr stark ausgelastet sind, oder die Zeitgarantien für Echtzeitanwendungen erfüllen sollen. Für geeignete Server ist die Konsolidierung empfehlenswert, da sie fast ausschließlich Vorteile birgt.

Der Einsatz von einem **zentralen Festplattenspeicher** im Netzwerk bietet durch die Reduktion der benötigten Speichermedien bei Verwendung größerer Kapazität ein weiteres großes Potential zur Leistungseinsparung. Auch hier ist eine Implementierung bis auf wenige Ausnahmen sinnvoll und kann zusätzlich noch die Datenübertragungsraten verbessern. Nicht geeignet sind Systeme, die exklusiven Zugriff auf die Datenträger benötigen.

Durch das Zusammenschalten kleinerer **Netzteile** können diese oftmals in effizienteren Lastbereichen betrieben werden. Ein interessanter Ansatz, der ohne fertige Produkte mit entsprechenden Spezifikationen und Garantien allerdings kaum empfohlen werden kann. Denkbar wären diese kombinierten Netzteile auch pro Rack, die je nach benötigter Leistung intern weitere Netzteile zuschalten können. Zudem kann man Lastspitzen durch die Versorgung mehrerer Server etwas glätten, was die Anforderungen bezüglich Leistungsschwankungen für die Netzteile reduziert.

Mit dem „**Spurt zum Schlafenlegen**“ können Server mit (sehr) kurzen Lastspitzen Leistung einsparen. Dabei wird der Prozessor mit möglichst hoher Frequenz betrieben, um früher wieder in einen (teil-)ausgeschalteten Zustand zurückkehren zu können. Ein Vorteil ist die reduzierte Latenz des angebotenen Dienstes, allerdings sollte über diese Maßnahme nur bei genauer Kenntnis über die Zahl und Dauer von Lastspitzen entschieden werden. Daher ist dieser Ansatz weniger für die Betreiber eines Rechenzentrums geeignet, als vielmehr für die Administratoren der Server. Eine Konsolidierung der Server ist aus Sicht des Rechenzentrum-Betreibers vermutlich sinnvoller.

Stark ausgelastete Server können ggf. auf mehrere Hosts **aufgeteilt** werden, um deren Prozessoren in niedrigeren Frequenzen betreiben zu können. Dieser Ansatz ist besonders bei VM-Hosts sinnvoll, bei denen durch die voneinander isolierten virtuellen Server keine Abhängigkeiten zwischen den Aufgaben bestehen. Auch andere gut skalierende Anwendungen können hiervon profitieren und dazu weniger stark unter Lastspitzen leiden. Für schlecht auf mehrere Systeme skalierbare Aufgaben ist dieser Ansatz jedoch gänzlich ungeeignet.

Eine **erhöhte Umgebungstemperatur** kann die Leistungsaufnahme der Lüfter und Klimatisierung reduzieren, womit diese anderweitig eingeplant werden kann. Bei Einhaltung der Spezifikationen für einen sicheren Betrieb der Server ist dieser Ansatz empfehlenswert. Jedoch kann es sinnvoll sein, den Fokus eher auf die aktive Kühlung in den Servern zu setzen. Langsam drehendere Lüfter können bei Lastspitzen viel schneller nachgeregelt werden, als die Klimaanlage die Lufttemperatur beeinflussen kann.

Mit der **Spannungs- und Frequenzskalierung** kann die Leistungsaufnahme pro Anwendung dynamisch reguliert werden. Auch hier liegt das Publikum eher bei den Serveradministratoren als bei den RZ-Betreibern, da Anwendungen Metriken bereitstellen müssen, um die geleistete Arbeit quantifizieren zu können. Dafür berücksichtigt dieser Ansatz auch wenig CPU-lastige Anwendungen und kann diese fair priorisieren. Zudem müssen Anwendungen auf bestimmten Kernen ausgeführt werden, deren Spannung und Frequenz von der Software kontrolliert wird.

Energie, die gerade nicht benötigt wird, kann in **Energiespeichern** vorgehalten werden und bei Bedarf Lastspitzen decken. Bei großen Lastwechseln kann dieser Ansatz helfen, eine größere Zahl an Servern zu versorgen ohne das Leistungsbudget zu überschreiten. Wichtig ist hierbei jedoch die Dimensionierung des Puffers, um genügend Energie vorhalten zu können.

6 FAZIT

Zusammenfassend lässt sich festhalten, dass durch die vorgestellten Ansätze große Möglichkeiten zur Einsparung von Energie existieren. Sowohl die Reduktion der Anzahl an physischen Servern, als auch die Optimierung der verbleibenden Maschinen ermöglichen es, einen Großteil der zuvor benötigten Leistung auf neue Systeme zu verteilen. Damit kann die Gesamtrechenleistung eines Rechenzentrums bei gleichem Leistungsbudget gesteigert werden, was sowohl der Profitabilität, als auch dem Umweltschutz zugute kommt.

LITERATUR

[1] L. A. Barroso and U. Hölzle. The Case for Energy-Proportional Computing. *Computer*, 40(12):33–37, 2007.

[2] S. Pelley, D. Meisner, T. F. Wenisch, and J. W. VanGilder. Understanding and Abstracting Total Data Center Power. 2009. http://web.eecs.umich.edu/~twenisch/papers/weed09.pdf (3.1.2021).

[3] N. Ahuja. Datacenter power savings through high ambient datacenter operation: CFD modeling study. In *2012 28th Annual IEEE Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM)*, pages 104–107, 2012.

[4] A. Guliani and M. M. Swift. Per-application power delivery. In *Proc. of the Fourteenth EuroSys Conference 2019*, pages 1–16, 2019.

[5] M. Gottschlag and F. Bellosa. Mechanism to Mitigate AVX-Induced Frequency Reduction, 2018. https://arxiv.org/abs/1901.04982 (3.1.2021).

[6] Jeffrey Stuecheli, Dimitris Kaseridis, Hillery C Hunter, and Lizy K John. Elastic refresh: Techniques to mitigate refresh penalties in high density memory. In *2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture*, pages 375–384. IEEE, 2010.

[7] R. Appuswamy, M. Olma, and A. Ailamaki. Scaling the Memory Power Wall With DRAM-Aware Data Management. In *Proc. of the 11th Int. Workshop on Data Management on New Hardware*. ACM, 2015.

[8] D. Meisner, B. T. Gold, and T. F. Wenisch. PowerNap: Eliminating Server Idle Power. In *Proc. of the 14th Int. Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS XIV, page 205–216. ACM, 2009.

[9] D. Li, J. S. Vetter, G. Marin, C. McCurdy, C. Cira, Z. Liu, and W. Yu. Identifying Opportunities for Byte-Addressable Non-Volatile Memory in Extreme-Scale Scientific Applications. In *2012 IEEE 26th Int. Parallel and Distributed Processing Symposium*, pages 945–956, 2012.

[10] J. H. Lee, M. H. Lee, and G. H. Jang. Experimental Verification of the Optimal FDBs in a HDD Spindle Motor to Minimize Power Loss. *IEEE Transactions on Magnetics*, 49(6):2437–2440, 2013.

[11] D. A. Patterson, G. Gibson, and R. H. Katz. A case for redundant arrays of inexpensive disks (RAID). In *Proc. of the 1988 ACM SIGMOD Int. conference on Management of data*, pages 109–116, 1988.

[12] E. Tomes and N. Altiparmak. A Comparative Study of HDD and SSD RAIDs' Impact on Server Energy Consumption. In *2017 IEEE Int. Conference on Cluster Computing (CLUSTER)*, pages 625–626, 2017.

[13] V. Kasavajhala. Solid state drive vs. hard disk drive price and performance study. *Proc. Dell Tech. White Paper*, pages 8–9, 2011.

[14] W. Huang, M. Allen-Ware, J. B. Carter, E. Elnozahy, H. Hamann, T. Keller, C. Lefurgy, Jian Li, K. Rajamani, and J. Rubio. TAPO: Thermal-aware power optimization techniques for servers and data centers. In *2011 Int. Green Computing Conference and Workshops*, 2011.

[15] M. Uddin and A. A. Rahman. Server consolidation: An approach to make data centers energy efficient and green. *arXiv preprint arXiv:1010.5037*, 2010.

[16] P. G. Jeba Leelipushpam and J. Sharmila. Live VM migration techniques in cloud environment — A survey. In *2013 IEEE Conference on Information Communication Technologies*, pages 408–413, 2013.

[17] B. Mao, H. Jiang, S. Wu, L. Tian, D. Feng, J. Chen, and L. Zeng. HPDA: A Hybrid Parity-Based Disk Array for Enhanced Performance and Reliability. *ACM Trans. Storage*, 8(1), February 2012.

[18] E. B. Nightingale, J. Elson, J. Fan, O. Hofmann, J. Howell, and Y. Suzue. Flat Datacenter Storage. In *10th USENIX Symposium on Operating Systems Design and Implementation (OSDI 12)*, pages 1–15, Hollywood, CA, October 2012. USENIX Association.

[19] B. Battles, C. Belleville, S. Grabau, and J. Maurier. Reducing Data Center Power Consumption Through Efficient Storage. 2007. https://storageconsortium.de/content/sites/default/files/downloads/wp-reducing-datacenter-power-consumption.pdf (3.1.2021).

[20] D. Wang, C. Ren, A. Sivasubramaniam, B. Urgaonkar, and H. Fathy. Energy Storage in Datacenters: What, Where, and How Much? *SIGMETRICS Perform. Eval. Rev.*, 40(1):187–198, June 2012.

[21] J. P. Barton and D. G. Infield. Energy storage and its use with intermittent renewable energy. *IEEE Transactions on Energy Conversion*, 19(2):441–448, 2004.