

## AWS Accelerated Computing auf Deutsch

### Was ist Accelerated Computing? (AWS Accelerated Computing)

In AWS bezeichnet **Accelerated Computing** EC2-Instanzen, die spezielle Hardware-Beschleuniger (wie **GPU, FPGA und AWS Inferentia**) verwenden, um bestimmte rechenintensive Workloads zu beschleunigen. Diese Beschleuniger ermöglichen eine deutlich schnellere Verarbeitung als herkömmliche CPU-basierte Berechnungen.

#### ✦ **Eigenschaften:**

- ✓ Nutzt spezielle Hardware (GPU, FPGA, ASIC), um rechenintensive Workloads zu beschleunigen
- ✓ Optimiert für Anwendungen wie **künstliche Intelligenz, maschinelles Lernen, Deep Learning und Videoverarbeitung**
- ✓ Ermöglicht parallele Verarbeitung und Datenparallelität für große Datensätze
- ✓ Ideal für Anwendungen mit hohen Leistungsanforderungen

---

### EC2-Instanztypen für Accelerated Computing

Die folgenden EC2-Instanztypen gehören zur Kategorie **Accelerated Computing** in AWS:

Instanztyp Eigenschaften		Anwendungsfälle
<b>P4d</b>	Nvidia A100 GPU, hohe Leistung für Deep Learning	Deep Learning, maschinelles Lernen
<b>G4dn</b>	Nvidia T4 GPU, Video-Encoding und Gaming-Server	Grafikverarbeitung, Video-Transcoding
<b>Inf1</b>	AWS Inferentia, optimiert für maschinelles Lernen	Modell-Inferenz, KI-Anwendungen
<b>F1</b>	FPGA-Beschleuniger, anpassbare Berechnungen	Hochleistungs-FPGA-Anwendungen
<b>P3</b>	Nvidia V100 GPU, Deep Learning und wissenschaftliches Rechnen	Deep Learning, Genomanalyse
<b>G5</b>	Nvidia A10G GPU, optimiert für Grafik- und Gaming-Workloads	Gaming, Grafikverarbeitung, AR/VR

💡 **P4d** und **P3** sind ideal für rechenintensive Workloads wie **Deep Learning und maschinelles Lernen**. **G4dn** und **G5** sind besser für **Grafik- und Videoverarbeitung** geeignet.

---

### Anwendungsfälle für Accelerated Computing

✅ **Accelerated Computing ist ideal für folgende Workloads:**

1 🤖 **Maschinelles Lernen und Deep Learning**

- ♦ Training von Modellen mit TensorFlow, PyTorch, MXNet

✅ **GPU- und Inferentia-Beschleuniger** verkürzen die Trainingszeit bei großen Datensätzen erheblich.

2 🎮 **Gaming und Grafikverarbeitung**

- ♦ Gaming-Server, Video-Transcoding und Rendering

✅ **T4- und A10G-GPUs** sorgen für eine schnelle Grafikverarbeitung.

3 🧬 **Wissenschaftliches Rechnen**

- ♦ Biologische Forschung, Genomanalyse, Simulationen

✅ **P3- und P4d-Instanzen** bieten mehr Rechenleistung für schnellere Analysen.

4 💡 **KI-Inferenz und Echtzeit-KI**

- ♦ Echtzeit-KI-Modelle, Sprach- und Bildverarbeitung

✅ **Inferentia-ASIC-Beschleuniger** ermöglichen eine effiziente KI-Inferenz.

5 ⚡ **Finanzdienstleistungen und Handel**

- ♦ Algorithmischer Handel, Risikoanalyse, Big Data-Verarbeitung

✅ **FPGA-Beschleuniger** bieten eine extrem schnelle Datenanalyse.


---

### Accelerated Computing vs. Andere Instanztypen

Eigenschaft	Accelerated Computing (P/G/F-Serie)	General Purpose (M-Serie)	Compute Optimized (C-Serie)
Einsatzbereich	GPU- und FPGA-Workloads, maschinelles Lernen	Ausgewogene CPU- und RAM-Nutzung	Hohe CPU-Lasten
Typische Workloads	Deep Learning, Grafikverarbeitung, FPGA-Anwendungen	Webserver, Unternehmensanwendungen	Gaming-Server, Videoverarbeitung
Beispielinstanzen	P4d, G4dn, Inf1, F1	M7i, M6i	C7g, C6i

---

### Solltest du Accelerated Computing nutzen?

 **Ja, wenn:**

- ✓ Du **hohe Rechenleistung und parallele Verarbeitung** benötigst (z. B. für Deep Learning, KI oder Grafikverarbeitung)
- ✓ Du Anwendungen für **maschinelles Lernen und künstliche Intelligenz** entwickelst
- ✓ Du **Grafik-Rendering, Video-Encoding oder Gaming-Server** betreiben möchtest.