NVIDIA Multi-Tenant AI Clouds antreiben: Warum Zadara die ideale Softwareplattform ist

Willkommen zu unserer neuen Blogserie, in der wir beleuchten, wie Zadara einzigartig positioniert ist, um die Software-Referenzarchitektur für Multi-Tenant-Inferenz-Clouds Realität werden zu lassen. Mit der nun veröffentlichten Blaupause von NVIDIA für eine mandantenfähige Generative-AI-Infrastruktur ist es an der Zeit, einen Blick darauf zu werfen, wie Cloud-Anbieter diese Vision in der Praxis umsetzen können. Wir bei Zadara sind überzeugt, dass wir Ihr idealer Partner dafür sind. In dieser Serie werden wir verschiedene Komponenten der Referenzarchitektur beleuchten – von GPU-Netzwerken bis hin zur Isolierung der Steuerungsebene – und zeigen, wie Zadara dies möglich macht.

Beginnen wir mit dem Gesamtbild: Was fordert NVIDIAs Referenzdesign – und warum ist Zadara bereits dafür gemacht?

Verständnis der Anforderungen der NVIDIA-Referenzarchitektur

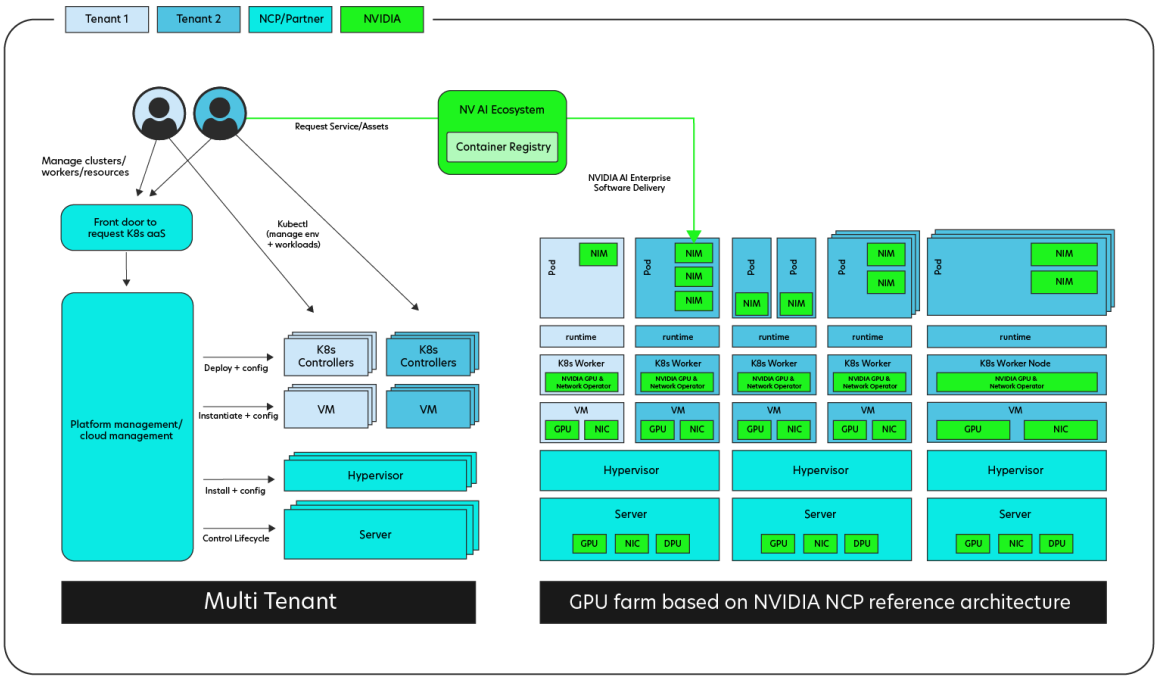
NVIDIAs Software-Referenzarchitektur ist ein umfassendes Framework, das Cloud-Service-Providern helfen soll, skalierbare, sichere und leistungsstarke KI-Infrastrukturen bereitzustellen. Im Kern unterstützt die Architektur:

- **Echte Mandantenfähigkeit**: Vollständige Isolation zwischen Kunden über die gesamte Technologiestack hinweg (Rechenleistung, Speicher, Netzwerk und Orchestrierung).
- **KI-zentrierte Infrastruktur**: Optimierung für KI-Workloads, die über GPU-Training hinausgehen – einschließlich Inferenz, Datenverarbeitung, Datenbanken und Orchestrierungsschichten.
- **Dynamische Ressourcenallokation**: Möglichkeit zur Bereitstellung und Skalierung von Ressourcen (GPUs, CPUs, Speicher, Netzwerk) pro Mandant und pro Workload.

- **Mandantenkontrollierte Kubernetes-Umgebungen**: Jeder Kunde sollte innerhalb seiner eigenen Kubernetes-Steuerungsebene arbeiten können, um maximale Flexibilität und Kontrolle zu gewährleisten.
- **Unterstützung für Edge- und Core-Bereitstellungen**: Die Architektur muss Bereitstellungen mit geringer Latenz nahe beim Nutzer ebenso wie zentralisierte Cloud-Operationen ermöglichen.

Da KI-Modelle zunehmend komplexer werden, steigt auch der Rechenaufwand für Inferenz – besonders bei Workloads, die auf logischem Schlussfolgern basieren, wie Entscheidungsbäume, Planung oder Codegenerierung. Solche Modelle benötigen häufig größere Speicherressourcen und längere GPU-Ausführungszeiten, wodurch dynamische, leistungsstarke Ressourcenallokation nicht nur fürs Training, sondern auch für Echtzeit-Inferenz unerlässlich wird.

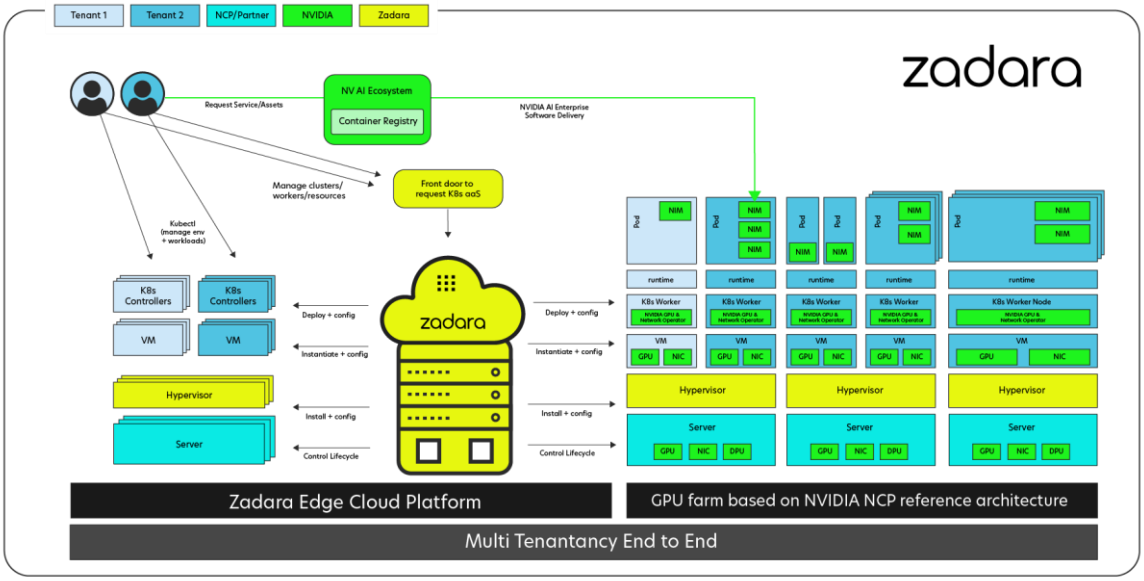
Diese Anforderungen werden durch NVIDIA-Hardware und -Software wie **Spectrum-X** (für Hochleistungsnetzwerke), **BlueField-3 DPU** (für ausgelagerte und sichere Netzwerke) und **NVIDIA AI Enterprise** (für KI-Betrieb) erfüllt.



Warum Zadara die richtige Wahl ist

Zadara wurde von Grund auf als Multi-Tenant-Cloud entwickelt und entspricht damit auf natürliche Weise den Empfehlungen von NVIDIA. Hier ein Überblick, wie Zadara die Anforderungen nicht nur erfüllt, sondern übertrifft:

- **Native Mandantenfähigkeit**:
  Zadara bietet integrierte Mandantenisolation für Rechenleistung, Speicher und Netzwerk. Jeder Mandant arbeitet in seinem eigenen sicheren Infrastruktursegment mit richtlinienbasierten Zugriffssteuerungen.
- **Full-Stack-Workload-Unterstützung**:
  Moderne KI-Workloads beschränken sich nicht nur auf GPUs. Zadara unterstützt alle Komponenten moderner KI/ML-Umgebungen – einschließlich Datenbanken, Vektor-Suchmaschinen und Kubernetes-Komponenten.
- **Dedizierte Kubernetes-Umgebungen pro Mandant**:
  Zadara ermöglicht die Bereitstellung separater Kubernetes-Steuerungsebenen für jeden Mandanten. Dies entspricht der Empfehlung zur Trennung der Steuerungsebenen und bietet höchste Flexibilität.
- **Elastische Ressourcenallokation**:
  Rechen-, Speicher- und GPU-Ressourcen können bei Zadara dynamisch pro Mandant und Workload zugewiesen werden. So wird die Infrastruktur effizient genutzt und flexibel skaliert.
- **Globale Edge-Präsenz**:
  Mit über 500 von Zadara betriebenen Edge-Standorten in mehr als 25 Ländern und über 200 regionalen Partnern bringt Zadara KI näher an die Nutzer. Dies ermöglicht Inferenz mit niedriger Latenz für verschiedenste Workloads – von RAG-basierten LLMs bis zu anspruchsvollen Aufgaben im logischen Denken. Gleichzeitig werden Anforderungen an Datenresidenz erfüllt – ein weiterer zentraler Punkt der Referenzarchitektur.



Wie geht’s weiter in der Serie?

In den nächsten Beiträgen gehen wir tiefer auf spezifische NVIDIA-Technologien ein und zeigen, wie Zadara diese unterstützt:

- **Spectrum-X und GPU-Netzwerke**: Aufbau der leistungsstarken Datenebene für KI.
- **BlueField DPUs**: Sichere und beschleunigte Netzwerke, schlanke Hypervisoren, Trennung von Steuerungs- und Laufzeitebene.
- **Isolation der Kubernetes-Steuerungsebene**: Wie Zadara die Orchestrierung pro Mandant im großen Maßstab ermöglicht.

Bleiben Sie dran – wir beleuchten bald jede dieser Schlüsselkomponenten. Die Zukunft der Multi-Tenant-AI-Clouds ist da – und sie läuft auf Zadara.
