memorysolution

Technisches Whitepaper

Cloud vs. On-Premise und Digitale Souveränität

Dieses Whitepaper richtet sich an IT-Leiter:innen, Systemarchitekt:innen und Infrastrukturverantwortliche in Unternehmen und öffentlichen Einrichtungen der DACH-Region – ebenso wie an Managed Service Provider, Systemhäuser, wissenschaftliche Einrichtungen und Forschungs- und Entwicklungsabteilungen großer Unternehmen. Es soll neutral, tiefgehend und praxisnah beleuchten, wie man im Jahr 2025 fundierte Architekturentscheidungen zwischen Cloud und On-Premises trifft – unter Berücksichtigung von Wirtschaftlichkeit, Sicherheit, Souveränität und technologischer Zukunftsfähigkeit. Dabei ist klar: Architekturentscheidungen lassen sich nicht pauschalisieren – sie müssen stets im Lichte der individuellen Anforderungen und Gegebenheiten getroffen werden.

Inhalt

1. Strategischer Kontext 2025:
Digitale Souveränität neu denken

a) Cloud-first war gestern – warum Architekturentscheidungen neu bewertet werden müssen

Das Jahr 2025 markiert einen Wendepunkt in der IT-Strategie vieler Unternehmen. Nach einer Dekade des „Cloud-first"-Paradigmas, rückt zunehmend die Frage in den Fokus, wie sich IT-Architekturen vor dem Hintergrund geopolitischer Spannungen, regulatorischer Anforderungen und des Anspruchs auf digitale Souveränität strategisch sinnvoll aufstellen lassen.

Insbesondere in der DACH-Region sind die Rahmenbedingungen komplex: Strenge Datenschutzgesetze (DSGVO), branchenspezifische Regulierungen (z. B. im Gesundheitswesen, in der Forschung oder im Finanzsektor) und ein wachsendes Bewusstsein für digitale Autonomie stellen technische Entscheider:innen vor neue Bewertungsmaßstäbe.

Hinzu kommt eine zunehmende Unsicherheit hinsichtlich transatlantischer Datenflüsse: Unter der US-Regierung von Donald Trump wurden gesetzliche Grundlagen wie der CLOUD Act aktiv durchgesetzt, wodurch US-Behörden weitreichende Zugriffsrechte auf in US-Clouds gespeicherte Daten erhielten – auch dann, wenn sich diese physisch auf Servern in Europa befinden. Trotz des 2023 in Kraft getretenen „EU-U.S. Data Privacy Framework" bleiben juristische Risiken und politische Unsicherheiten bestehen, die den Druck auf Unternehmen in sensiblen Sektoren weiter erhöhen, ihre IT-Strategien souverän und resilient auszurichten.

b) Repatriierung & Realitätscheck: Warum die Cloud an Grenzen stößt

Gleichzeitig haben Cloud-Angebote in den letzten Jahren enorme Fortschritte in puncto Funktionalität, Skalierbarkeit und globaler Verfügbarkeit gemacht. Dennoch ist das Thema 2025 strategisch relevanter denn je: Viele Organisationen mussten in der Praxis feststellen, dass Cloud-Lösungen nicht per se kostengünstiger oder einfacher sind – insbesondere bei dauerhaft hohen Lasten, sensiblen Daten oder komplexen Integrationsanforderungen.

Steigende Betriebskosten, Sicherheitsbedenken und Performance-Probleme führen zunehmend dazu, dass Unternehmen produktive Workloads aus der Public Cloud zurück in eigene Rechenzentren verlagern – ein Trend, der unter dem Begriff „Cloud Repatriierung" längst keine Randerscheinung mehr ist. Gerade in datenintensiven und regulierten Branchen zeigt sich: **Die Cloud ist nicht mehr alternativlos, sondern muss im Kontext von Wirtschaftlichkeit, Kontrolle und regulatorischer Passung neu bewertet werden.**

c) Digitale Souveränität als betriebliche Notwendigkeit

Zum anderen zwingen neue EU-Regularien und Initiativen (z. B. GAIA-X, EU Digital Services Act) dazu, genauer hinzuschauen, wo und wie Daten verarbeitet werden. Digitale Souveränität ist vom Buzzword zur konkreten Anforderung geworden: Eine aktuelle EU-Umfrage ergab, dass 72 % der europäischen Unternehmen Datensouveränität als entscheidenden Faktor bei der Auswahl von Technologieanbietern nennen – gegenüber 58 % im Jahr 2022 (Hivenet).

Entsprechend steigt die Nachfrage nach IT-Modellen, die Datenhoheit, Zugriffs- und Betriebskontrolle gewährleisten. Neben souveränen europäischen Cloud-Angeboten wächst auch das Interesse an On-Premise-Infrastrukturen und Private-Cloud-Architekturen, die sich nahtlos in hybride Betriebsmodelle integrieren lassen. Besonders Organisationen mit sensiblen oder geschäftskritischen Daten prüfen zunehmend, welche Workloads dauerhaft lokal betrieben werden sollten – nicht zuletzt, um regulatorische Klarheit, technische Kontrolle und wirtschaftliche Planbarkeit zu vereinen.

Kriterium	Cloud (Public)	Eigene Infrastruktur
Kosten bei Dauerlast	🔴 Hoch	🟢 Planbar
Datenschutz (DSGVO)	🟠 Abhängig	🟢 Volle Kontrolle
Digitale Souveränität	🟠 Eingeschränkt	🟢 Hoch
Skalierbarkeit	🟢 Sehr hoch	🟠 Planungsintensiv
Integration	🟠 Abhängig	🟢 Anpassbar

d) Über dieses Whitepaper

Dieses Whitepaper richtet sich an IT-Leiter:innen, Systemarchitekt:innen und Infrastrukturverantwortliche in Unternehmen und öffentlichen Einrichtungen der DACH-Region – ebenso wie an Managed Service Provider, Systemhäuser, wissenschaftliche Einrichtungen und Forschungs- und Entwicklungsabteilungen großer Unternehmen. Es soll neutral, tiefgehend und praxisnah beleuchten, wie man im Jahr 2025 fundierte Architekturentscheidungen zwischen Cloud und On-Premises trifft – unter Berücksichtigung von Wirtschaftlichkeit, Sicherheit, Souveränität und technologischer Zukunftsfähigkeit.

Wir betrachten wirtschaftliche Aspekte (TCO, Betriebskosten, FinOps), technische Überlegungen (Leistung, Workload-Platzierung, Integration) sowie Sicherheits- und Souveränitätsfragen (Compliance, Zertifizierungen, souveräne Cloud-Stacks). Reale Projektbeispiele – anonymisiert oder fiktiv an realen Anforderungen orientiert – zeigen Hürden, Lösungen und Lessons Learned.

Wichtig ist ein nüchterner, faktenbasierter Vergleich der Vor- und Nachteile: Weder „Alles muss in die Cloud" noch „Cloud ist Teufelszeug" bringen eine zukunftsfähige IT-Strategie voran. Stattdessen gilt es, für jeden Anwendungsfall das passende Betriebsmodell zu finden und – wo sinnvoll – Cloud- und On-Prem-Welten zu kombinieren.

Dieses Whitepaper wurde von der Memorysolution GmbH verfasst, einem spezialisierten Distributor für Server-Infrastrukturen, Enterprise-Storage und individuelle Hardwarelösungen. Gemeinsam mit unseren Partnern haben wir bereits zahlreiche maßgeschneiderte Private-Cloud-Umgebungen On-Premises realisiert – insbesondere für Kunden mit besonderen Anforderungen an Datenhoheit, Performance und Integrationstiefe. Dennoch verfolgen wir in diesem Whitepaper das Ziel, eine so unvoreingenommene wie fundierte Entscheidungsgrundlage zu schaffen – mit konkretem Praxisbezug, strategischer Orientierung und klaren Handlungsempfehlungen bis hin zu Checklisten für die Umsetzung.

Jetzt beraten lassen

Memorysolution GmbH

Memorysolution ist seit 1997 auf hochwertige Speicherlösungen, Serverkomponenten und IT-Infrastrukturen spezialisiert. Als Distributor mit starken Herstellerpartnerschaften – u. a. mit Samsung, Supermicro und **KIOXIA** – bietet das Unternehmen seit 2016 auch individuelle Server- und Storagelösungen unter dem Label Mustang Systems an.

2. Strategische Grundlagen:
Cloud, On-Premises, Hybrid & Edge

Bevor wir ins Detail gehen, lohnt ein präziser Überblick zentraler Betriebsmodelle und aktueller Trends:

a) On-Premises

- IT-Infrastruktur (Server, Speicher, Netzwerk) wird im eigenen Rechenzentrum oder direkt vor Ort beim Kunden betrieben
- Volle Kontrolle über Hardware, Software, Netzwerke und Sicherheitsarchitektur
- Trotz Cloud-Boom weiterhin stark verbreitet
- Gründe: Datenhoheit, regulatorische Anforderungen, bestehende Investitionen, nicht cloudfähige Legacy-Systeme

b) Public Cloud

- IT-Ressourcen (Compute, Storage, Plattformdienste etc.) aus Rechenzentren großer Anbieter wie AWS, Microsoft Azure oder Google Cloud
- Zugriff über das öffentliche Internet, Abrechnung meist nutzungsbasiert („pay as you go“)
- Vorteile: hohe Skalierbarkeit, globale Verfügbarkeit, kein eigener Hardware-Betrieb notwendig
- Nachteile: begrenzte Kontrolle, Lock-in-Risiken, laufende Kosten schwer kalkulierbar, Abhängigkeit vom Anbieter

c) Private Cloud

- Anwendung von Cloud-Prinzipien (Self-Service, Elastizität, Automatisierung) auf dedizierte Infrastruktur
- Betrieb im eigenen RZ oder bei einem vertrauenswürdigen Hosting-/Managed-Service-Partner
- Vorteile: hohe Kontrolle, bessere Datenschutzoptionen, flexibler Betrieb auf eigener Hardware
- Nachteile: Verantwortung für Infrastrukturmanagement verbleibt beim Betreiber (intern oder extern beauftragt)

94 % der Unternehmen berichten von zumindest etwas Cloud-Waste, 49 % schätzen, dass mehr als 25 % ihrer Public-Cloud-Ausgaben verschwendet werden.
Quelle: Private Cloud Outlook 2025 Report von Broadcom

66 % der Entwickler gehen davon aus, dass > 20 % ihrer Cloud-Infrastrukturkosten verschwendet werden – Hauptursachen sind unterausgelastete Ressourcen und fehlende Sichtbarkeit.
Quelle: „FinOps in Focus 2025“ von Harness

Bevor wir ins Detail gehen, lohnt ein präziser Überblick zentraler Betriebsmodelle und aktueller Trends:

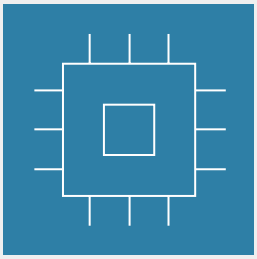
d) Hybrid Cloud

- Kombination aus On-Premises- und Cloud-Ressourcen mit abgestimmter Integration
- 72 % der Unternehmen weltweit setzen auf Hybrid 2024
  Quelle: IDC / Flexera (State of the Cloud Report 2025)
- Typisch: sensible Daten bleiben lokal, skalierbare Dienste (z. B. AI-Training, Backups, Peak-Load) laufen in der Cloud
- Herausforderungen: Orchestrierung, Datenabgleich, Sicherheitskonzepte, Latenzmanagement

e) Multi-Cloud

- Nutzung mehrerer Public-Cloud-Anbieter (z. B. AWS und Azure) parallel
- Ziel: Reduktion von Anbieterabhängigkeiten, Nutzung spezialisierter Services je Plattform
- Erfordert hohes technisches Know-how und durchdachtes Plattformmanagement

f) Edge Computing



- Verlagerung von Rechenleistung an den Rand des Netzwerks – möglichst nah an der Datenquelle oder dem Endanwender
- Typisch: kompakte, energieeffiziente Systeme direkt vor Ort (z. B. in Produktionshallen, Retail-Filialen, Fahrzeugen)
- Einsatz bei latenzkritischen Anwendungen und dezentraler Datenverarbeitung in Echtzeit
- Häufig als Teil hybrider Architekturen, synchronisiert mit zentralen Systemen (Cloud oder RZ)

Im Jahr 2023 entfielen bereits 31 % der IT-Budgets auf Hybrid Cloud – und IT-Leitende erwarten, dass dieser Anteil in den nächsten zwei Jahren auf 48 % steigt.
Quelle: Data Centre Review – „The next chapter of cloud in Europe"

83 % der CIOs planen Cloud-Repatriierung – teilweise Rückverlagerung von Workloads zurück in On-Prem oder Private Cloud, globaler Trend sichtbar.
Quelle: Barclays CIO Survey, 1H 2024

3. Aktuelle Entwicklungen und Bewertungskriterien

a) Trends 2025: Konvergenz von Cloud und On-Premise

Die Grenzen zwischen Cloud und On-Premise verwischen zusehends. Große Cloud-Anbieter investieren verstärkt in Technologien, die einheitliches Management über hybride Umgebungen ermöglichen:

Azure Arc

von Microsoft erlaubt die zentrale Verwaltung von Ressourcen, egal ob diese in Azure, On-Premises oder in anderen Clouds laufen. VMs, Kubernetes-Cluster oder Datenbanken können über Azure-Dienste wie Policies, Monitoring und Security-Services verwaltet werden – auch außerhalb der Microsoft-Cloud.

AWS Outposts

bringt native AWS-Infrastruktur ins eigene Rechenzentrum. Dabei handelt es sich um vollständig verwaltete Rack-Systeme, die lokal installiert werden und dieselben APIs, Tools und Services wie die AWS-Regionen nutzen – mit extrem niedriger Latenz und voller Kontrolle.

Google Anthos

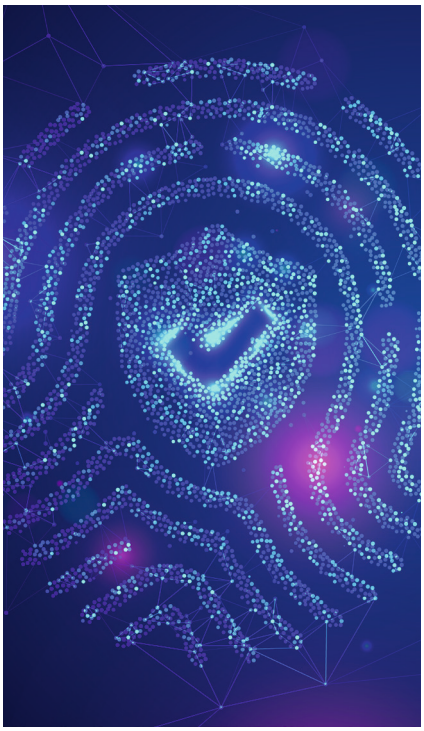
ist eine Plattform für den Betrieb und das Management containerisierter Anwendungen über mehrere Umgebungen hinweg – sowohl in GCP als auch On-Premise oder bei anderen Cloud-Anbietern. Grundlage ist ein Kubernetes-zentriertes Modell mit zentraler Governance, CI/CD, Service Mesh und Policy Enforcement.

Auch On-Premises-Technologien entwickeln sich weiter, um moderne Cloud-Betriebsmodelle bereitzustellen. Container-Orchestrierung mit Kubernetes, Automatisierung via Ansible, Software-defined Storage oder OpenStack ermöglichen skalierbare, selbstverwaltbare Infrastruktur mit Cloud-ähnlicher Agilität – im eigenen RZ.

Parallel entstehen zunehmend Angebote im Bereich **Sovereign Cloud:**

> **Hyperscaler kooperieren mit lokalen Partnern oder schaffen abgeschottete Regionen, um regulatorische Anforderungen an Datenhoheit und Betreiberverantwortung zu erfüllen.**

So hat **AWS 2023 seine „European Sovereign Cloud"** für den öffentlichen Sektor angekündigt. Microsoft führt eine **EU Data Boundary** ein und arbeitet mit regionalen Partnern an Cloud-Diensten, die ausschließlich von EU-Personal in europäischen Rechenzentren betrieben werden.



b) Architektursicht: Die richtigen Fragen stellen

Es gibt keine universelle Architekturentscheidung –
sinnvoll ist, jede IT-Landschaft anhand zentraler Leitfragen zu bewerten:

Compliance & Datenhoheit

Gibt es rechtliche oder interne Vorgaben, wo Daten verarbeitet oder gespeichert werden dürfen? (z. B. DSGVO, Geheimschutz, Branchenregeln)

Performance & Latenz

Welche Workloads erfordern lokale Verarbeitung, hohe Bandbreite oder minimale Reaktionszeiten?

Skalierbarkeit & Lastprofil

Sind die Anforderungen konstant oder schwanken sie stark?
Wann lohnt sich Elastizität – und wann ist eine hohe, stabile Auslastung wirtschaftlicher?

Bereitstellungszeit

Wie schnell müssen Systeme produktiv sein –
Stunden, Tage oder Monate?

Bestehende Infrastruktur & Know-how

Gibt es ein RZ und entsprechendes Betriebspersonal,
oder starten wir auf der grünen Wiese?

Finanzierungsmodell

Passt CAPEX (On-Prem) oder OPEX (Cloud) besser zur Strategie?

Diese Fragen sind die Grundlage für fundierte Architekturentscheidungen –
im nächsten Abschnitt betrachten wir die wirtschaftlichen und betrieblichen Aspekte im Detail.

4. Wirtschaftlichkeit & Betrieb
im Vergleich

a) TCO, Repatriierung und FinOps

Eine Kernfrage ist die Wirtschaftlichkeit von Cloud vs. On-Premise über mehrere Jahre. Oft wird argumentiert, Cloud spare Kosten, da keine hohen Anfangsinvestitionen anfallen und man nur zahlt, was man nutzt. Doch diese Rechnung geht nicht in jedem Szenario auf. Entscheidend ist eine Total Cost of Ownership (TCO)-Betrachtung über den Lebenszyklus (typisch 3–5 Jahre), die alle Kostenfaktoren einbezieht:

Negativ:
Unausgelastete Ressourcen (z.B. Server laufen auf 20 % CPU-Last) stellen gebundenes Kapital dar.

Positiv:
Die Kosten sind relativ planbar und fix, unabhängig von der Nutzung – was Überaschungen minimiert.

Kostenfaktoren On-Premises:

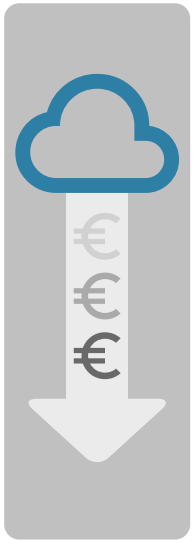
- Hardwarekosten für Server, Storage, Netzwerkkomponenten
- Flächenkosten: Miete oder Abschreibung für RZ-Gebäude/-Fläche
- Energieverbrauch: Stromversorgung und Kühlung
- Wartung und Upgrades: Ersatzteile, Wartungsverträge, Modernisierung
- Betriebskosten für Personal (Admin, Support, Monitoring)
- Softwarelizenzen: z. B. VMware, Datenbank-Software, Backup-Lösungen
- Supportverträge mit Dienstleistern

Kostenverteilung:

- Teils upfront (CAPEX, z. B. Hardware)
- Teils laufend (OPEX, z. B. Strom, Personal, Wartung)
- Typischer Nebeneffekt: Geringe Auslastung
  (z. B. Server oft <30 % CPU), führt zu ineffizienter Kapitalbindung.

Negativ:
Die Cloud-Rechnung setzt sich aus vielen Posten zusammen. Egress-Kosten für ausgehende Datenmengen können unerwartet teuer werden.

Positiv:
Hier gibt es kaum Anfangsinvestition – man zahlt monatlich für tatsächlich genutzte Ressourcen.



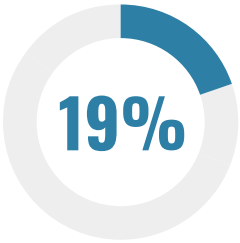
Kostenfaktoren Cloud (OPEX-basiert):

- Compute-Kosten (VMs, Container, Autoscaling-Instanzen)
- Speicherkosten (z. B. pro GB Block-, Objekt- oder Dateispeicher)
- Netzwerkkosten, insb. für ausgehende Daten (Egress Fees)
- Managed Services: Datenbanken, Backup, Monitoring, etc.
- API-/Transaktionskosten (z. B. bei hohem Zugriff auf Microservices)
- Nutzungslastabhängige Preisgestaltung (Skalierung je nach Traffic)
- Regionale Preisunterschiede je nach Standort der Cloud-Ressourcen
- Zusatzkosten durch ungenutzte Ressourcen („Cloud-Waste“)
- Kostenrisiko: Hohe Variabilität, schwer planbar, Budgetüberschreitungen häufig.

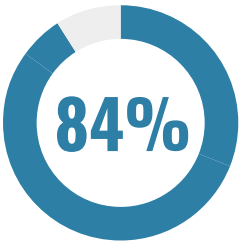
b) Volatile Kosten in der Cloud – und warum FinOps zur Pflicht wird

i.Wirtschaftliche Aspekte: Cloud-Kosten unter Druck

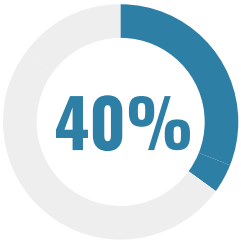
Cloud-Kosten sind variabel und skalieren mit der Nutzung. Im Idealfall können Unternehmen bei geringer Last Ressourcen reduzieren und somit Kosten sparen. In der Praxis treten jedoch häufig Ineffizienzen auf: Nicht abgeschaltete Instanzen, überdimensionierte Ressourcen oder vergessene Testumgebungen verursachen sogenannten „Cloud-Waste“. Eine aktuelle Studie (FinOps in Focus 2025 von Harness) kommt auf einen durchschnittlichen Anteil von 21 % verschwendeter Cloud-Ausgaben. Der Private Cloud Outlook 2025 Report von Broadcom zeigt zudem, dass 49 % der befragten IT-Entscheider schätzen, mehr als ein Viertel ihrer Public-Cloud-Ausgaben werde verschwendet. In Branchenberichten und Marktanalysen ist darüber hinaus häufig von 30 % oder mehr die Rede.



Canalys (Cloud Service Spending 2025) prognostiziert für 2025 einen weltweiten Anstieg der Cloud-Kosten um 19 %, hauptsächlich getrieben durch den zunehmenden Einsatz von KI-Diensten.



Flexera (State of the Cloud Report 2025) berichtet, dass 84 % der Unternehmen Schwierigkeiten haben, ihre Cloud-Ausgaben zu kontrollieren. Ein Drittel der Unternehmen gibt jährlich über 12 Millionen US-Dollar für Public-Cloud-Dienste aus, was die Budgets um durchschnittlich 17 % übersteigt.



Microsoft hat angekündigt, die Preise für Cloud-Dienste wie Microsoft 365, Teams und Azure ab April 2025 um bis zu 40 % zu erhöhen. Dies ist bereits die dritte signifikante Preiserhöhung innerhalb weniger Jahre.

Diese Entwicklungen unterstreichen die Notwendigkeit eines effektiven Cloud-Kostenmanagements. Ansätze wie FinOps gewinnen an Bedeutung, um Transparenz zu schaffen und Ausgaben zu optimieren.

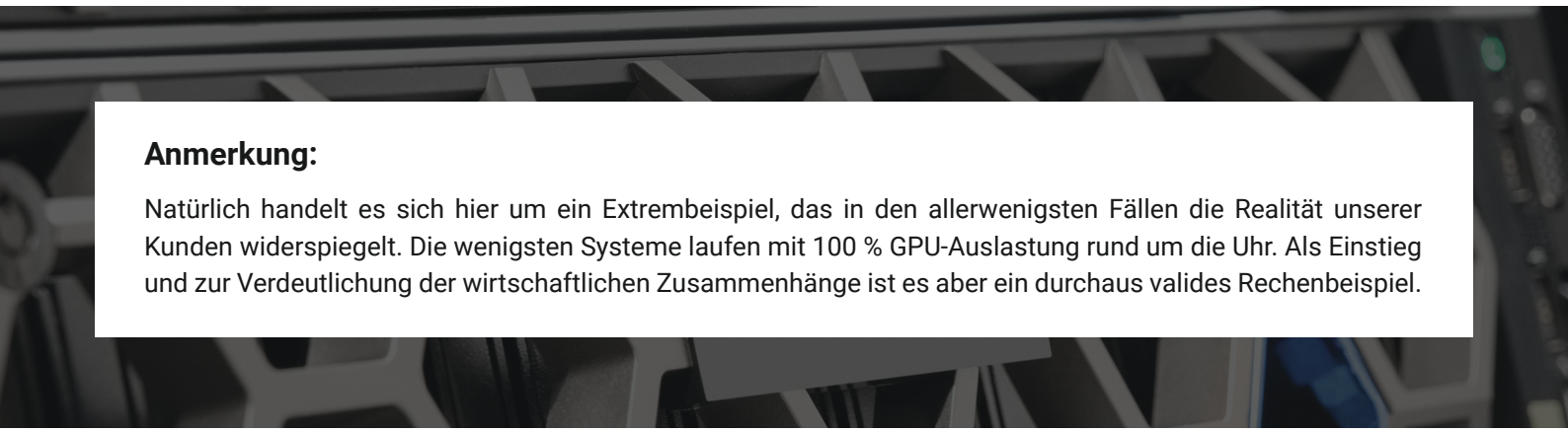
Ohne aktive Kostenkontrolle können Cloud-Ausgaben schnell aus dem Ruder laufen. Viele CIOs haben erkannt, dass Cloud-Kostenmanagement – oft unter dem Begriff FinOps – mittlerweile zur Pflichtdisziplin gehört. Über die Hälfte der Unternehmen nannte 2024 die Optimierung von Cloud-Ausgaben und die Vermeidung unnötiger Ressourcenverschwendung als oberste Priorität.

c) On-Prem vs. Cloud: TCO im direkten Vergleich

i.TCO-Vergleich über 3–5 Jahre

Pauschal lässt sich nicht sagen, welches Modell günstiger ist – entscheidend ist das konkrete Nutzungsverhalten. Bei dauerhaft hoher Auslastung kann On-Premise erhebliche wirtschaftliche Vorteile bieten, da Investitionen über mehrere Jahre hinweg abgeschrieben werden, während Cloud-Dienste dauerhaft laufende Kosten erzeugen.
Ein anschauliches, wenn auch zugespitztes **Beispiel liefert eine TCO-Analyse von Lenovo für KI-Hardware:**

- Ein GPU-basierter Server verursacht On-Premise über fünf Jahre Gesamtkosten von rund 872.000 US-Dollar (inkl. Strom und Betrieb).
- Die vergleichbare Nutzung über AWS p5 Instances kommt im selben Zeitraum auf über 4,3 Mio. US-Dollar.
- Selbst mit Rabattmodellen wie 3-Jahres-Reserved-Instances bleibt die Cloud deutlich teurer – in diesem Fall rund 1,5 Mio. USD Differenz.
- Der Break-Even wird bereits nach etwa 12 Monaten erreicht – ab diesem Punkt spart jede weitere Betriebsstunde On-Premise bares Geld.



Anmerkung:

Natürlich handelt es sich hier um ein Extrembeispiel, das in den allerwenigsten Fällen die Realität unserer Kunden widerspiegelt. Die wenigsten Systeme laufen mit 100 % GPU-Auslastung rund um die Uhr. Als Einstieg und zur Verdeutlichung der wirtschaftlichen Zusammenhänge ist es aber ein durchaus valides Rechenbeispiel.

ii. Wirtschaftliche Daumenregeln: Wann lohnt sich On-Premise?

Bei geringerer Auslastung verschiebt sich der Break-Even: Läuft ein System z.B. nur ~5 Stunden pro Tag, ist Cloud meist günstiger. Die oben zitierte Analyse ermittelte für ihr Beispiel eine Schwelle von ~5 Stunden/Tag – darüber lohnt sich on-prem, darunter die Cloud. Mit Rabatten (Reserved/Savings Plans) steigt die Schwelle auf ~6–9 Stunden/Tag.

iii. FinOps trifft On-Prem: Zwei Welten, zwei Hebel

Hier kommt das Konzept FinOps ins Spiel – Financial Operations für die Cloud. FinOps bedeutet, Cloud-Ressourcen und -Ausgaben laufend zu überwachen, Budgets einzuhalten und die Kosten pro Nutzungseinheit zu optimieren.
On-Prem hat dagegen fixe Kapazitäten – die Herausforderung ist hier, diese möglichst gut auszulasten. Im Grunde braucht es auch on-prem ein Kostenbewusstsein, allerdings mit anderen Hebeln.

Grundsätzlich gilt:

- Je kontinuierlicher und vorhersagbarer ein Workload, desto eher kann on-prem ökonomisch sinnvoll sein.
- Kurzlebige, volatile Workloads spielen ihre Stärken in der Cloud aus.
- Wird eine Nutzung jedoch dauerhaft und planbar, wachsen die laufenden Cloud-Kosten erheblich.
- In solchen Fällen amortisiert sich eine Investition in eigene Hardware oft nach wenigen Jahren oder sogar Monaten.

memorysolution

Wirtschaftlich rechnen. Strategisch planen. Passend Konfigurieren.

Mit individuell zusammengestellten Serversystemen schaffen Sie die Basis für souveräne, wirtschaftlich betreibbare IT-Infrastrukturen.

- Maßgeschneiderte Serverlösungen auf Supermicro-Basis
- Optimiert für TCO, Performance und Zukunftssicherheit
- Ideal für Virtualisierung, HPC und KI-Workloads

Sprechen Sie mit uns über Ihre Anforderungen

5. Sicherheit & Souveränität: Anbieter, Anforderungen, Standards, Risiken

a) Digitale Souveränität im Fokus

Neben Kosten und Performance rücken Security, Compliance und vor allem die Frage der digitalen Souveränität immer stärker in den Vordergrund – besonders in der DACH-Region mit ihren hohen Datenschutzanforderungen und dem zunehmenden Wunsch nach Unabhängigkeit von außereuropäischen Tech-Giganten.

Doch was bedeutet digitale Souveränität operativ?
Und wie unterscheiden sich Cloud- und On-Prem-Modelle in diesem Kontext?

Datensouveränität bedeutet, dass eine Organisation die volle Kontrolle darüber hat, wo und wie ihre digitalen Daten gespeichert und verarbeitet werden, wer darauf zugreifen kann – und welchem Rechtsraum sie unterliegen. In der Praxis geht es z. B. darum, ob personenbezogene oder sensible Daten das EU-Gebiet verlassen oder ob ausländische Behörden – auch ohne konkreten Anlass – Zugriff darauf erzwingen könnten.

Einordnung: Was ist der CLOUD Act (USA)?
Der 2018 in Kraft getretene „Clarifying Lawful Overseas Use of Data Act" verpflichtet US-amerikanische Technologieunternehmen dazu, auf richterliche Anordnung hin Daten offenzulegen – auch wenn sich diese Daten außerhalb der USA befinden, etwa in Rechenzentren in Europa. Das bedeutet: Selbst wenn ein US-Cloud-Anbieter Server in Frankfurt oder Zürich betreibt, könnten US-Behörden (z. B. FBI, NSA) im Rahmen bestimmter Verfahren Zugriff auf dort gespeicherte Daten verlangen – ohne dass europäische Datenschutzstandards dem wirksam entgegenstehen.

Für Unternehmen mit hohen Anforderungen an Vertraulichkeit, Geheimschutz oder DSGVO-Konformität ist das ein ernst zu nehmendes Risiko – das bei der Wahl der Infrastruktur berücksichtigt werden muss.

b) On-Premises: Maximale Kontrolle, maximale Verantwortung

On-Premises bietet hier naturgemäß maximale Kontrolle: Daten verbleiben im eigenen Haus/Rechenzentrum, und man kann technisch wie organisatorisch sicherstellen, dass kein Dritter Zugang hat (sofern die eigenen Sicherheitsmaßnahmen robust sind). Allerdings muss man dafür die Verantwortung komplett selbst tragen – Security ist on-prem eigenverantwortlich.

Es bedarf Expertise, um etwa Netzwerke segmentiert abzusichern, Firewalls korrekt zu konfigurieren, regelmäßige Patches einzuspielen und Monitoring/Incident Response aufzubauen.
Die Cloud nimmt einem gewisse Sicherheitsaufgaben ab (etwa physische Sicherheit im Rechenzentrum, Grundschutz der Infrastruktur), doch bleibt viel in Kundenhand (Zugriffskonzepte, Konfiguration etc.). Shared Responsibility lautet das Modell – und Fehlkonfigurationen in der Cloud führen leider häufig zu Sicherheitsvorfällen (offene S3-Buckets etc.).

Ein Vorteil von On-Prem: Man kann kundenspezifische Security-Maßnahmen* tiefergehend umsetzen und ist nicht auf die Tools des Providers beschränkt . Beispielsweise lassen sich spezielle Verschlüsselungstechniken oder Netzwerkappliances integrieren, was in einer Public Cloud-Umgebung ggf. nicht geht.

c) Public Cloud Security & Compliance: Standardzertifizierungen vs. Rechtsrahmen

Große Cloud-Provider investieren massiv in Security und erreichen i.d.R. ein sehr hohes Grundniveau an Schutz (DDOS-Schutz, professionelle Physische Sicherheit, oft automatische Verschlüsselung von Daten at-rest, etc.). Sie bieten zahlreiche Compliance-Zertifizierungen out-of-the-box:

ISO 27001
Internationaler Standard für Informationssicherheits-Managementsysteme (ISMS).

SOC 2
Prüft, ob Cloud-Dienste definierte Sicherheits- und Datenschutzkriterien einhalten.

PCI-DSS
Sicherheitsstandard für den Umgang mit Kreditkartendaten.

HIPAA
US-Gesetz zur Absicherung medizinischer Daten und Patientenschutz.

BSI C5
Deutscher Mindeststandard für sichere Cloud-Anbieter (Cloud Computing Compliance Controls Catalogue).

Wer also z.B. eine ISO 27001 konforme Umgebung braucht, findet in der Cloud bereits zertifizierte Services vor. Gerade für kleinere Firmen kann das attraktiv sein, weil sie nicht alles selbst auditieren lassen müssen – sie nutzen die Zertifikate des Providers mit.

Aber: Die Compliance mit Datenschutzgesetzen wie DSGVO erfordert dennoch vertragliche und technische Maßnahmen (Auftragsverarbeitung, ggf. EU-Standardvertragsklauseln).

Ein Knackpunkt bleibt die Frage der jurisdiktionalen Kontrolle: Wenn ein Cloud-Anbieter aus den USA stammt (auch wenn Rechenzentren in EU stehen), gibt es die Sorge, dass US-Gesetze (wie CLOUD Act) den Zugriff auf Daten erlauben könnten, ohne dass europäische Stellen dem zustimmen.

Diese Unsicherheit führt dazu, dass viele Behörden und regulierte Branchen Cloud-Dienste von US-Anbietern skeptisch sehen, solange keine rechtliche Absicherung besteht. Microsoft hat z.B. reagiert und 2023 die Initiative ergriffen, alle EU-Kundendaten ausschließlich in der EU zu verarbeiten („EU Data Boundary“) und zuzusichern, dass Zugriffe nur nach EU-Recht erfolgen. Ob dies im Konfliktfall hält, wird sich zeigen.

d) Initiativen zur Stärkung souveräner Cloud-Angebote

i. GAIA-X – Föderiertes Cloud-Ökosystem nach europäischen Werten

In Europa laufen mehrere Initiativen, um digitale Souveränität zu stärken. GAIA-X ist eine vielbeachtete davon – ein Verbundprojekt, um ein **föderiertes, sicheres Cloud- und Datenökosystem** nach europäischen Werten aufzubauen.

Dabei geht es weniger um einen einzelnen Cloud-Service, sondern um Standards und Zertifizierungen.

GAIA-X definiert etwa „Label" für Cloud-Services (Level 1–3), wobei die höchste Stufe (Level 3) sicherstellen soll, dass:

> **Daten ausschließlich in Europa verarbeitet werden und**
>
> **keinerlei ausländischer Einfluss besteht.**

Ein ganz praktischer Schritt in 2025: Erste Anbieter erhalten GAIA-X-konforme Labels, und Tools wie der Cloud Data Engine ermöglichen es, automatisch zu prüfen, ob ein Service die geforderten Souveränitätskriterien einhält. So kann z.B. ein Datenmarktplatz sicherstellen, dass nur Cloud- Dienste eingebunden werden, die garantiert EU-only sind (kein Transfer in Drittstaaten, Betreiber unter EU-Recht etc.).

ii. EU-Politik und neue Regulierungsansätze

Auch die EU-Kommission selbst forciert das Thema: Margrethe Vestager wurde 2024 zur ersten EU Kommissarin für Technologie-Souveränität ernannt, um Cloud-Regularien und faire Wettbewerbsbedingungen durchzusetzen.

Neue Gesetzespakete wie **Digital Markets Act (DMA)** und **Digital Services Act (DSA)** zielen zwar primär auf Plattformregulierung ab, aber flankieren die Souveränitätsbestrebungen, indem Big-Tech in Schranken gewiesen und Transparenzpflichten erhöht werden.

Speziell für Cloud kommt voraussichtlich der **EU Cloud Rulebook** und ein EU Zertifizierungsschema (EUCS), das Cloud-Diensten je nach Erfüllung von Souveränitätskriterien ein Gütesiegel geben soll . **Allerdings wird noch diskutiert, ob EUCS in der höchsten Stufe eine Besitzstruktur in der EU und Immunität gegen Nicht-EU-Recht voraussetzt – was US-Hyperscaler ausschließen würde, es sei denn, sie gründen rechtlich unabhängige Ableger.**

iii. Lokale Anbieter und Open-Source-Stacks: Bausteine für Souveränität

Für Organisationen, die aus regulatorischen oder strategischen Gründen keine US-Clouds nutzen dürfen oder wollen – etwa weil Forschungsdaten nicht abfließen dürfen oder nationale Sicherheitsvorgaben greifen – gibt es in der DACH-Region eine Vielzahl von Alternativen:

OpenTelekomCloud
der Deutschen Telekom

IONOS
(Deutschland)

Swisscom Cloud
(Schweiz)

sowie zahlreiche regionale Rechenzentrumsbetreiber mit Datentreuhandmodellen

Auch einige Hyperscaler versuchen, über Partnerschaften „souveräne" Angebote zu schaffen. So hat **Google gemeinsam mit T-Systems** eine souveräne Cloud-Lösung für Deutschland angekündigt, bei der T-Systems als Datentreuhänder fungiert. Frühere Modelle wie **Azure Deutschland mit Telekom-Treuhänderschaf**t wurden jedoch wieder eingestellt – was die Herausforderungen solcher Konstrukte unterstreicht.
Auf Software-Ebene setzen viele souveräne Anbieter und öffentliche Einrichtungen auf **Open-Source-Technologien** wie:

- Nextcloud (File Collaboration)
- Matrix (dezentrale Chat- und Kommunikationsplattform)
- OpenStack (Infrastructure-as-a-Service)
- Kubernetes (Orchestrierung containerisierter Anwendungen)
- Ceph und DAOS (verteilte Storage-Infrastrukturen)

hosttech – Sichere Infrastrukturen aus der Schweiz

hosttech ist ein führender Schweizer Anbieter von Hosting-, Domain-, Cloud- und Managed-Services mit Hauptsitz in Richterswil (ZH) sowie Standorten in Deutschland und Österreich. Das Unternehmen betreibt eigene, ISO-27001-zertifizierte Rechenzentren – darunter das hochsichere Tier-3-Bunker-Datacenter DATAROCK in Nottwil. hosttech bietet seinen Kunden 100 % Datenhaltung in der Schweiz oder innerhalb der DACH-Region, vollständig konform mit DSGVO und frei von außereuropäischen Rechtszugriffen.

- Hosting & Cloud „Made in Switzerland“
- ISO-27001-zertifizierte Rechenzentren (u. a. Tier-3-Bunker DATAROCK)
- Volle DSGVO- und Compliance-Sicherheitn
- Flexible Standortwahl (Schweiz, Deutschland, Österreich) für Virtual Datacente
- Langjährige Erfahrung in Hosting, Domains, vServer und Cloud-Infrastrukturen

„Die Zusammenarbeit mit dem Memorysolution-Team ist herausragend. Die Geschwindigkeit, mit der Anfragen und Offerten bearbeitet werden, ist beeindruckend. Dazu kommt die 1A-Qualität bei jeder Bestellung – von der sorgfältigen Verkabelung bis zur sicheren Verpackung.“

Marius Meuwly
CEO hosttech GmbH

hosttech

Über hosttech

hosttech betreibt moderne, ISO-27001-zertifizierte Rechenzentren in der Schweiz, Deutschland und Österreich – darunter das hochsichere Tier-3-Bunker-Datacenter DATAROCK in Nottwil. Seit 2004 bietet hosttech zuverlässige Hosting-, Domain-, Cloud- und Managed-Services. Mit Fokus auf Sicherheit, Datensouveränität und Performance unterstützt hosttech Unternehmen jeder Größe – von KMU bis hin zu Organisationen in regulierten Branchen.

Memorysolution

Ihr Partner für maßgeschneiderte Server- & Storage-Infrastrukturen.

Mehr zu Memorysolution

hosttech

Sichere Hosting-, Cloud- und Managed-Services aus der Schweiz.

Cloud-Services ansehen

Projekt-Highlight:

croit GmbH – Souveräner Ceph-/DAOS-Stack mit Memorysolution

Projektsteckbrief

Projekt:
Ceph- & DAOS-Storage-Infrastruktur

Kunde:
croit GmbH

Branche:
Software-defined Storage (SDS)

Projektzeitraum / Volumen:
2020–2024 / ca. 2.700.000 €

Hardwarekonzept & Technologien

Ceph-Cluster:

- Supermicro All-Flash-Server
- AMD EPYC CPUs
- Seagate Mach.2 SAS-HDDs

DAOS-Cluster:

- Supermicro-Server mit Intel Xeon Scalable Gen 3
- Intel Optane Persistent Memory
- Hochparallele I/O mit minimaler Latenz

Storage-Performance:

- NVMe-Speicher mit bis zu 30,72 TB U.3-NVMe-Drives pro Server
- Über 6 TB/s Bandbreite in IO500-Benchmarks – weltweite Spitzenklasse

Zur vollständigen Projektbeschreibung

Hintergrund: Open-Source-SDS für maximale Unabhängigkeit

Die **croit GmbH** entwickelt skalierbare, flexible und einfach bedienbare Speicherlösungen auf Basis von **Ceph** und **DAOS** – zwei leistungsstarken Open-Source-Technologien im Bereich softwaredefinierter Speicherinfrastrukturen. Die Trennung von Hardware und Software ermöglicht höchste **Unabhängigkeit**, eine flexible Weiterentwicklung und **Investitionssicherheit ohne Herstellerbindung.**

Projektschwerpunkte

- Aufbau einer **hochperformanten, selbstheilenden, hochverfügbaren Ceph/DAOS-Infrastruktur**
- Realisiert mit **über 200 individuell konfigurierten Serversystemen** von Mustang Systems (Memorysolution)
- Ergänzt durch **120 weitere Systeme** für Co-Location-Projekte und Hochschulpartnerschaften
- Performanceziel: Bandbreitenklasse der Weltspitze – mit Erfolg

Warum Ceph & DAOS von croit?

- Herstellerneutral & vendor-lock-in-frei
- Offene Architektur & vollständige Transparenz
- Kostenkontrolle & Skalierbarkeit durch Open Source
- 24/7 Support & professioneller Betrieb
- Schulungsangebote & Know-how-Transfer inklusive
- Perfekte Basis für sensitive, regulatorisch anspruchsvolle Datenumgebungen

Rolle von Memorysolution

Als Technikpartner haben wir das Projekt über den gesamten Zyklus hinweg begleitet:

Auswahl & Feinkonzeption der Hardware

Teststellungen & Performance-Validierung

Rack-Integration & Rollout

After-Sales-Support & Erweiterungen für Universitätsprojekte

Das Projekt zeigt exemplarisch, wie **maßgeschneiderte Hardwarelösungen** mit **Open-Source-Softwarestacks** kombinierbar sind – und wie **technologische Souveränität** auch bei extremen Performanceanforderungen **keine Kompromisse** machen muss.

iv. Offene Stacks als Souveränitätsvorteil

Diese Stack-Entscheidungen fließen in Souveränitätsbewertungen ein – Quelloffenheit und Auditierbarkeit sind Pluspunkte. GAIA-X selbst setzt auf offene Schnittstellen und **Interoperabilität** als Prinzip, um einen **vendor lock-in** zu vermeiden. **RZ-Standards und Zertifizierungen:** In der DACH-Region sind gewisse Standards fast Pflicht, um als „vertrauenswürdig" zu gelten – sei es in der Cloud oder on-prem.

v. Infrastruktur-Standards im DACH-Raum

Weit verbreitete Standards und Zertifizierungen

ISO/IEC 27001 — Informationssicherheits-Managementsystem

Viele Unternehmen zertifizieren ihr eigenes Rechenzentrum oder verlangen dies von Cloud-Anbietern, um ein Grundniveau an Security-Prozessen nachzuweisen.

BSI C5 — Cloud Computing Compliance Criteria Catalogue

Ein deutscher Anforderungskatalog speziell für Cloud-Dienste, der vom BSI veröffentlicht wird. Cloud-Provider (inkl. AWS, Azure, IBM etc.) lassen regelmäßig C5-Testate erstellen. Für deutsche Behörden und den öffentlichen Sektor ist ein C5-Nachweis quasi obligatorisch, um einen Cloud-Service zu nutzen. C5 kombiniert internationale Standards mit zusätzlichen Anforderungen für Transparenz und Sicherheit aus deutscher Sicht.

IT-Grundschutz — BSI

Ein umfassender Katalog an Sicherheitsmaßnahmen, der auch als Zertifizierung erworben werden kann. Öffentliche Stellen in DE orientieren sich daran. Besonders kritische Infrastrukturen (KRITIS) müssen Auflagen aus BSI-Gesetzen erfüllen.

Branchenspezifische Standards:	
TISAX	für Automobilindustrie-Zulieferer, Fokus Datensicherheit
PCI-DSS	für Kreditkartendaten, falls relevant
HIPAA	im Gesundheitsbereich, v.a. für US aber auch für globale Pharma-Unternehmen relevant) oder spezifische Aufsichtsanforderungen wie BAIT/VAIT (für Banken/Versicherer IT in Deutschland

Wer Cloud nutzt, muss sicherstellen, dass der Anbieter diese Anforderungen unterstützt (viele haben entsprechende Compliance-Berichte).

Rechenzentrums-Zertifizierungen:

TÜV- oder EN 50600-Auditierung für Planung, Bau und Betrieb von Rechenzentren – relevant für physische Sicherheit, Verfügbarkeit und betriebliche Zuverlässigkeit.

Uptime Institute TIER-Zertifizierungen (Tier III / IV) geben Auskunft über das angestrebte Verfügbarkeitslevel. Zwar nicht direkt ein Souveränitätskriterium, aber entscheidend für die Vertrauenswürdigkeit der Infrastruktur: Ein hochsouveränes System nützt wenig, wenn das Rechenzentrum selbst unsicher oder störanfällig ist.

vi. Operative Souveränität: Mehr als Standortwahl

Es reicht nicht, die Daten physisch in der EU liegen zu haben. Operative digitale Souveränität bedeutet deutlich mehr:

> **Nur autorisiertes Personal mit entsprechender Sicherheitsfreigabe darf auf Systeme zugreifen**
>
> **Alle Verwaltungsfunktionen werden vom europäischen Ableger aus geführt**
>
> **Logs, Systeminformationen und Metadaten verbleiben vollständig in Europa**

Viele Cloud-Anbieter betonen inzwischen „EU resident support and operations" – also: Der operative Betrieb erfolgt durch in der EU ansässiges Personal.

Beispiele aktueller Cloud-Initiativen

Microsoft kündigte an, Kunden die Wahl zu geben, dass ausschließlich EU-Betriebspersonal für deren Tenant zuständig ist.

Google Cloud und **Oracle** betreiben Cloud-Regionen in der EU mit der Zusicherung, dass weder Daten noch Metadaten außerhalb Europas verarbeitet werden – und sie unterwerfen sich europäischen Prüfmechanismen.

Diese Entwicklungen zeigen: **Hyperscaler reagieren auf den politischen Druck** – insbesondere aus der öffentlichen Hand.

vii. Fallbeispiel: KI-Modelle und Forschungsdaten außerhalb der US-Cloud?

Gerade beim Einsatz von KI-Technologien zeigen sich konkrete Risiken in Bezug auf Souveränität:
Wenn z. B. ein Krankenhaus sensitive Patientendaten für eine KI-Diagnose nutzen will, darf es diese kaum in eine US-Cloud laden – selbst anonymisierte Gesundheitsdaten gelten häufig als zu kritisch.

Zulässige und realistische Alternativen:

Für sicherheitskritische Forschungsbereiche – z. B. Rüstungsforschung oder verschlusssacherelevante Projekte – ist On-Premises heute ohnehin Standard.

- **On-Premise-Lösungen** oder **europäische Cloud-Infrastrukturen**
- **European Health Data Space (EHDS):** Die neue Verordnung schreibt vor, dass Gesundheitsdaten **in vertrauenswürdigen, europäischen Umgebungen** geteilt und verarbeitet werden müssen
- **Gaia-X** wird explizit als Infrastrukturbaustein für EHDS genannt

Technische Gegenmaßnahmen bei Cloud-Nutzung:

In weniger sensiblen Fällen lassen sich technische Maßnahmen einsetzen, um Souveränität trotz Cloud-Nutzung zu gewährleisten:

- **Client-seitige Verschlüsselung** (Schlüssel verbleiben on-prem)
- **Homomorphic Encryption**
- **Secure Multi-Party Computation**
- **Confidential Computing** (z. B. Trusted Execution Environments direkt auf Cloud-CPUs)

All diese Technologien befinden sich jedoch noch in früher Marktreife. In der Praxis wird meist der konservativere Weg gewählt: Daten bleiben lokal.

viii. Cloud-KI-Dienste: Souveränität oder Komfort?

Wenn KI-Dienste wie etwa GPT-Modelle via Cloud-API genutzt werden sollen, stellt sich zwangsläufig die Frage:
Darf ich meine Eingabedaten (Prompts) überhaupt extern verarbeiten lassen?
Viele Organisationen – insbesondere im Gesundheits-, Forschungs- oder Industrieumfeld – verbieten dies intern aus Gründen des Geheimnisschutzes oder der DSGVO-Konformität.

Gegenbewegung: Souveräne KI-Modelle

- Europäische LLMs (z. B. Aleph Alpha)
- Open-Source-Modelle, die on-prem betrieben werden
- Erkenntnisse von Dell Technologies zeigen: Das Inferieren großer KI-Modelle on-prem kann bis zu 75 % kosteneffizienter sein als über Public-Cloud-APIs

Neben der Kostenersparnis spricht ein weiteres Argument für On-Prem-KI: Die Kontrolle über Trainings-, Input- und Metadaten bleibt vollständig erhalten.

ix. Fazit: Souveränität operativ leben

Digitale Souveränität erfordert eine bewusste Auswahl der Infrastruktur – abgestimmt auf rechtliche, wirtschaftliche und sicherheitstechnische Rahmenbedingungen.

> **On-Premises bietet maximale Hoheit über Daten und Systeme – setzt aber entsprechende Ressourcen und Know-how voraus.**
>
> **Public Clouds punkten mit professioneller Infrastruktur, hoher Verfügbarkeit und zahlreichen Zertifizierungen – bergen aber Rechts- und Abhängigkeitsrisiken, die aktiv gemanagt werden müssen.**

Handlungsfelder

- Verschlüsselungstechnologien und Zugriffskontrolle
- Klare vertragliche Regelungen (z. B. DPA, AVV)
- Multi-Cloud-Strategien zur Reduktion von Lock-in
- Regelmäßige Audits & Compliance-Checks

In sensiblen Bereichen wie Gesundheitsdaten oder Forschung mit Sicherheitsauflagen bleibt On-Prem oder eine souveräne europäische Cloud meist alternativlos.

6. Technologische Umsetzung: Architektur, Infrastruktur, Integration

Nachdem wir strategisch das Für und Wider beleuchtet haben, stellt sich die Frage:

- Wie setzt man eine moderne, zukunftssichere On-Premise-Infrastruktur konkret um?
- Und wie integriert man sie bei Bedarf sinnvoll mit Public-Cloud-Services?
- Welche Technologien und Konzepte sind 2025 relevant, um das Beste aus beiden Welten zu verbinden?

a) Die eigene Infrastruktur „cloudy“ machen

Moderne On-Prem-Umgebungen sollen möglichst selbstbedienungsfähig, automatisiert und flexibel sein – also funktional an eine Cloud erinnern. Ziel ist der Aufbau einer Private Cloud oder zumindest einer „Cloud-nahen“ Betriebsumgebung.

Virtualisierung & Containerisierung

- VMs (z. B. VMware vSphere, Microsoft Hyper-V, KVM, Proxmox) sind Standard
- Container-Technologie (Docker, Kubernetes) gewinnt an Bedeutung
- Kubernetes als universelle Plattform für portable, cloud-native Workloads
- Hybride Kubernetes-Ansätze wie Red Hat OpenShift oder SUSE Rancher kombinieren On-Prem und Cloud
- Vorteil: Container ermöglichen Microservices-Architekturen, Skalierung & Automatisierung auch On-Premises

Orchestrierung & Self-Service:

Eine Private Cloud sollte es Entwicklern und Fachabteilungen ermöglichen, selbstständig Ressourcen wie VMs oder Storage bereitzustellen – ohne jedes Mal Tickets an die IT zu schreiben. Dafür braucht es Orchestrierungs-Tools und Self-Service-Portale.
Open-Source-Lösungen wie OpenStack bieten ein Dashboard und APIs ähnlich AWS (für VM, Netzwerk, Storage). Kommerzielle Alternativen sind z. B. VMware vRealize Automation oder CloudStack. Auch Infrastructure-as-Code ist on-prem möglich: Mit Tools wie Terraform oder Ansible lassen sich Ressourcen definieren und automatisiert bereitstellen – als internes IaaS-Modell.

Automation und DevOps-Praktiken:

Durchgehende Automatisierung (per Skripte, Ansible- Playbooks, CI/CD-Pipelines für Infrastruktur) ist entscheidend, um eine On-Prem-Umgebung effizient zu betreiben.
Viele greifen bewährte DevOps-Tools aus der Cloud-Welt auf:

- Git für Versionskontrolle von Konfigs
- Jenkins/GitLab CI für Deployment-Pipelines
- Monitoring mit Prometheus/Grafana
- Logging mit ELK-Stack etc

Das Ziel ist, die Betriebseffizienz so hoch wie möglich zu halten, damit On-Prem nicht personell ausufert. Einige Unternehmen investieren in Self-Service-Portale mit Katalogen (VM-Vorlagen, Datenbank-as-a-Service intern, Kubernetes-Namensräume etc.), um den Entwicklern Cloud-ähnliche Geschwindigkeit intern zu bieten.

Sicherheit und Identity-Management integrieren:
Ein Vorteil von On-Prem-Lösungen ist die nahtlose Anbindung an bestehende Authentifizierungssysteme wie Active Directory. Moderne Ansätze setzen auf Identity Federation, um Single Sign-On zwischen Cloud und On-Prem zu ermöglichen – etwa via Azure AD Connect, das lokale AD-Konten mit der Azure Cloud synchronisiert. So können Mitarbeitende beide Welten mit denselben Credentials nutzen. Zugriffskonzepte sollten konsistent aufgebaut sein: Prinzip der geringsten Rechte, flächendeckende MFA, Zero Trust und Netzwerksegmentierung. Häufig genutzte Lösungen sind zentralisierte LDAP/AD-Dienste, OAuth2/SSO für Applikationen sowie Cloud-Security-Tools, die auch On-Prem-Workloads erfassen.

b) Projekt: Hochperformanter Virtualisierungs-Cluster für ein Medienunternehmen

Projektzeitraum: 2024
Projektvolumen: ca. 900.000 €

Kunde:

Ein international tätiges Medienhaus mit hohem Bedarf an leistungsfähiger, virtualisierter IT-Infrastruktur.

Ausgangssituation:

Gesucht wurde eine hochverfügbare Plattform zur Virtualisierung von rund 200 Servern – inklusive schneller NVMe-Speicherlösung für geschäftskritische Anwendungen, großvolumigem HDD-Archiv für Langzeitdaten sowie maximaler Skalierbarkeit für zukünftiges Wachstum. Ein nahtloser Anschluss an die bestehende Systemlandschaft war essenziell.

Umsetzung durch Memorysolution mit Mustang Systems:

- Virtualisierungsplattform:
  - 3-Knoten-Proxmox-Cluster
- Compute Nodes:
  - 3× Supermicro A+ Server 2125HS-TNR
  - Jeweils 2× AMD EPYC™ 9554 (64 Cores)
  - 1.536 GB DDR5 RAM pro Node
  - 10× 7,68 TB NVMe pro Node
- Storage-Anbindung:
  - Broadcom 9500-8e HBAs
- - 3× 44-Bay JBODs (Supermicro 847E1C-R1K23JBOD)
  - Gesamt: 230 TB NVMe-Storage + 1,5 PB HDD-Archiv
- Netzwerk:
  - Redundanter 100 GBit/s-Backbone

Ergebnis:

- Performante Virtualisierungsplattform für 200 produktive Workloads
- Ultra-schneller NVMe-Zugriff für latenzkritische Anwendungen
- Petabyte-Speichervolumen für Langzeitdaten
- Zukunftssichere Architektur mit hoher Skalierbarkeit
- Redundante, hochverfügbare Netzwerkstruktur

Fazit:

Die schlüsselfertig übergebene Lösung integriert sich nahtlos in die bestehende IT-Umgebung und erfüllt höchste Anforderungen an Performance, Verfügbarkeit und Flexibilität. Das Projekt belegt eindrucksvoll die Kompetenz von Memorysolution und Mustang Systems im Bereich maßgeschneiderter High-Performance-Cluster – von der Hardware über das Architekturdesign bis hin zur operativen Beratung.

Jetzt beraten lassen

c) Server- und Storage-Konzepte 2025 für hybride Szenarien

Die Hardware-Landschaft entwickelt sich ständig weiter.

Aktuelle Trends, die in On-Prem-Architekturen eine Rolle spielen:

Hyperkonvergente Systeme (HCI):

Hier werden Compute und Storage in einem Cluster vereint – statt großer separater Storage-Arrays nutzt man die internen Disks/SSDs der Server, über Software zu einem Verbund gekoppelt (z.B. VMware vSAN, Nutanix, Microsoft S2D).

HCI-Systeme sind skalierbar durch einfaches Hinzufügen von Knoten, was Cloud-ähnliche Skalierung onprem erlaubt. Sie eignen sich gut für private Clouds bis mittlere Größe und vereinfachen den Betrieb (eine einheitliche Plattform für VM und Storage). Viele HCI-Lösungen bieten zudem Integrationen Richtung Cloud (Backup oder Tiering von Daten in Cloud-Speicher).

All-Flash-Storage & NVMe:

Die Preise für Flash-Speicher sind gefallen, und die Leistungsanforderungen gestiegen – 2025 sind All-Flash-Arrays oder zumindest NVMe-basierte Storage-Lösungen Stand der Technik für performante Workloads (Datenbanken, VDI, Analytics).

HDDs werden noch für Kapazität/Archiv genutzt, aber im Primärspeicher dominieren SSDs wegen geringer Latenz. NVMe-oF (NVMe over Fabrics) erlaubt es, ultraschnellen Flash-Speicher übers Netzwerk verfügbar zu machen mit kaum mehr Latenz als lokal. Unternehmen, die hohe IOPS brauchen, setzen z.B. NVMe Flash Arrays mit 100 Gbit iSCSI/NVMeoF ein.

GPU-Server und Beschleuniger:

Durch den KI-Boom sind GPU-Beschleuniger mittlerweile in vielen Bereichen gefragt – sei es Machine Learning, Bilderkennung, Simulation oder auch Virtual Desktop Infra für 3D-Anwendungen.

On-Premise kann man dedizierte GPU-Server einplanen (z.B. mit NVIDIA A100/H100 oder ähnlichen), eventuell in einem separaten KI-Cluster. Diese Hardware ist teuer, aber kontinuierliche Nutzung kann den Kauf rechtfertigen (siehe Kostenbeispiel oben). Alternativ bieten Clouds GPUs on-demand zu hohen Stundensätzen.

2025 kommen neben GPUs auch AI-ASICs (spezialisierte KI-Chips) und FPGAs vereinzelt zum Einsatz – hier sollte man die Marktentwicklung beobachten, etwa Graphcore IPUs oder Google TPUs (die aber nur in Google-Cloud verfügbar sind). Für Edge-Standorte gibt es kompakte Server mit kleineren GPUs (oder neu: inferenzoptimierte CPUs wie Intel Xeon mit AMX oder ARM-basierte mit NPUs).

Kompakte, energieeffiziente Systeme:

Da Energiepreise hoch sind und Platz in kleinen RZs begrenzt, achten viele auf dichte Formfaktoren und Effizienz. Heute erhält man z.B. 1-HE-Server mit 2×64 Kern CPUs und 1 TB RAM – eine Rechenleistung, die vor wenigen Jahren ein halbes Rack brauchte.

Blade-Systeme und modulare Microserver (z.B. 8 Nodes in 2HE Gehäuse) ermöglichen es, viel Compute auf wenig Raum unterzubringen, was auch die Kühlung pro Fläche vereinfacht.

Zudem setzen manche Organisationen auf ARM-basierte Server (wie Ampere Altra CPUs), die pro Core weniger Strom verbrauchen – für Cloud-nativen Workload können solche ARM-Server effizienter sein. Wichtig für 2025 ist auch das Thema Kühlung: Hochleistungs-GPUs erzeugen große Abwärme, weshalb in HPC- und KI-Clustern verstärkt auf Direkt-Flüssigkühlung gesetzt wird. Das erhöht Effizienz und könnte für nachhaltigkeitsbewusste Unternehmen ein Aspekt sein.

Multi-Cluster und verteilte Standorte:

In hybriden Szenarien betreibt man ggf. mehrere Standorte – etwa zentrale Rechenzentren und Edge-Installationen. Technologien wie Software- Defined WAN (SD-WAN) helfen, diese Standorte sicher und performant zu vernetzen.

Zudem wird oft eine zentrale Management-Ebene angestrebt, um verteilte Ressourcen zu verwalten (z.B. ein zentrales vCenter für mehrere ESXi-Standorte, oder Kubernetes Federation für mehrere Cluster).

Hier gilt es, Latenzen und Ausfallszenarien zu bedenken: Edge-Knoten sollten auch autonom weiterlaufen können, falls die Verbindung zur Zentrale mal wegfällt (Stichwort: Lokaler Controller/Fallback-Modus).

Backup, DR und Datenhaltung:

On-Premises muss man sich selbst um Backups kümmern - was aber auch mehr Freiheit gibt, wie und wo man sichert. Viele nutzen Hybrid-Cloud-Backup: Die laufenden Systeme on-prem, aber ein Backup-Repository wird verschlüsselt in Cloud-Speicher (z.B. AWS S3 oder Azure Blob) ausgelagert, um Offsite-Kopien zu haben.

Das verbindet die Vorteile – schnelle lokale Recovery-Möglichkeit plus geographische Redundanz. Disaster Recovery as a Service gibt es auch: Anbieter ermöglichen VM-Replikation ins Cloud- Rechenzentrum, sodass man im Notfall dort hochfahren kann. Andersherum kann man Cloud-Daten on-prem sichern (etwa Cloud-to-Cloud-Backup-Anbieter oder eigene Lösungen via APIs), wenn man vermeiden will, dass ein Cloud-Ausfall alle Daten unzugänglich macht.

d) Cloud- und On-Prem-Welten verbinden

Die Integration beider Welten zählt zu den größten technischen Herausforderungen hybrider Architekturen – aber auch zu den Bereichen mit ausgereiften Best Practices.

Netzwerkanbindung & Latenz

Für echten Hybridbetrieb sind dedizierte Verbindungen zur Cloud (z. B. VPN, Direct Connect, ExpressRoute) unerlässlich – sie bieten höhere Bandbreite und geringere Latenz als öffentliche Internetpfade.
Trotzdem liegen Latenzen selbst bei Direktverbindungen oft bei 10–30 ms. Latenzempfindliche Komponenten (z. B. DB und App-Server mit vielen synchronen Transaktionen) sollten daher möglichst lokal zusammengehalten werden – alternativ helfen Caching oder Replikation.

Datenabgleich & Synchronisation

Hybride Szenarien erfordern oft, Daten parallel in Cloud und On-Prem vorzuhalten. Replikationstools wie Oracle Data Guard, SQL Server Always On oder dateibasierte Ansätze (z. B. rsync, Cloud-Storage-Gateways) sind hier verbreitet.
Streaming-Daten lassen sich via Apache Kafka über Standorte hinweg synchronisieren. Wichtig sind ausreichend Bandbreite, klare Konsistenzmodelle und technisches Konfliktmanagement.
In regulierten Szenarien bleibt das „System of Record" meist On-Prem – Cloud dient nur der Analyse oder Aggregation.

Anwendungsintegration

Für hybride Applikationen braucht es klar definierte Schnittstellen – idealerweise über interne APIs. API-Gateways oder Service Meshes (lokal & cloudfähig) ermöglichen Steuerung, Security und Entkopplung.
Ein typisches Setup: Ein Web-Frontend läuft in der Cloud, ruft via VPN eine On-Prem-API auf. Wichtig sind sichere Authentifizierung (z. B. JWT, OAuth), Verschlüsselung, ggf. Caching oder asynchrone Aufrufe zur Latenzreduktion.

Zugriffskonzepte & Identity Management

Mitarbeitende sollen unabhängig vom Ort konsistenten Zugriff haben – zentrale Identitätslösungen (Azure AD, LDAP, SSO) sind daher essenziell.
Zugriffsrechte müssen klar geregelt sein – zunehmend setzen Unternehmen auf Cloud-Center-of-Excellence-Teams mit Domänen-übergreifendem Know-how.
Zentrales Security-Monitoring (z. B. via Sentinel, Splunk) hilft, Silostrukturen zu vermeiden.

Unified Monitoring & Logging

Einheitliche Überwachung steigert Effizienz und Fehlertoleranz. Tools wie Prometheus, Grafana oder der Elastic Stack können sowohl Cloud- als auch On-Prem-Systeme erfassen.
Alternativ integrieren Cloud-Plattformen On-Prem-Ressourcen über eigene Tools wie Azure Arc oder AWS Systems Manager.

Workload-Platzierung

Verursacht der Cloud-Workload hohe Egress-Kosten oder ständige Datentransfers ins eigene Rechenzentrum, kann eine Repatriierung wirtschaftlich oder technisch (Stichwort: Latenzen) sinnvoll. In vielen Fällen ist es effizienter, den Workload näher an der Datenquelle zu betreiben, etwa im eigenen Rechenzentrum.

Edge-Integration

Edge-Computing ergänzt hybride Architekturen: Lokale Edge-Knoten (z. B. in Produktionsanlagen) erfassen Daten und synchronisieren diese mit zentraler Cloud-Infrastruktur.
Sie müssen autonom funktionsfähig bleiben und robust gegenüber Netzwerkausfällen sein (lokale Persistenz, asynchrone Verarbeitung).
Sicherheitsaspekte: Keine eingehenden Verbindungen, Segmentierung, nur initiierte Außenkommunikation. Auch hier sind kompakte, energieeffiziente Systeme gefragt (z. B. Mini-Server für Schaltschrankmontage).

Fazit:

Hybride Infrastrukturen sind technisch anspruchsvoll, aber mit geeigneter Planung, standardisierten Protokollen (z. B. REST, OpenID, Kubernetes) und skalierbaren Tools gut umsetzbar.
Wichtig: Insellösungen vermeiden und Architektur flexibel halten – für echte Zukunftsfähigkeit ohne Anbieter- oder Technologie-Lock-in.

7. Projektbeispiele & Lessons Learned

Theorie ist gut – doch wie sehen solche Überlegungen in der Praxis aus? In diesem Kapitel betrachten wir reale bzw. realitätsnahe Projektbeispiele aus der DACH-Region, die ähnliche Anforderungen und Entscheidungen durchlaufen haben.

Die Beispiele sind anonymisiert oder fiktiv auf Basis echter Erfahrungen, um konkrete Lessons Learned aufzuzeigen. Zudem beleuchten wir, welche Rolle Memorysolution und Partner wie Menzel IT bei solchen Projekten spielen können, um maßgeschneiderte Lösungen zu liefern.

Zur vollständigen Success Story

a) Projektbeispiel 1: Hochskalierbare AMD-EPYC-Cloudplattform für Energie- und Industriekunden

Ausgangssituation

Ein führender deutscher IT-Dienstleister für Energie- und Industriekunden betreibt komplexe Infrastrukturen mit höchsten Anforderungen an Verfügbarkeit, Sicherheit und Skalierbarkeit. Um Wachstum abzufedern und Compliance-Vorgaben einzuhalten, sollte am Hauptstandort eine neue Cloud-Service-Plattform entstehen – leistungsfähig, souverän und erweiterbar für künftige Standorte.

Umsetzung in der Praxis

Die gewählte Lösung basiert vollständig auf **AMD EPYC 9004-Prozessoren** und einer **All-Flash-Architektur** mit Samsung NVMe-Laufwerken. Damit wurde eine hochperformante On-Premises-Plattform realisiert, die den Anforderungen an Sicherheit, Skalierbarkeit und Datenkontrolle gerecht wird.

Zentrale Komponenten:

- **10× Supermicro Hyper A+ 1125HS-TNR** mit je 2× AMD EPYC 9534 (64 Kerne) und 1,152 TB DDR5-4800 ECC RAM als Plattform-Rückgrat
- **3× Manager Nodes** mit AMD EPYC 9124 für Orchestrierung und Verwaltung
- **3× Control Nodes** mit AMD EPYC 9334 für Steuerung und Monitoring
- **All-Flash-Storage:** Samsung PM9A3 (OS) und PM1743 (Daten) mit PLP und AES-256 Hardwareverschlüsselung
- **Dual-100 GbE QSFP28** mit PCIe 5.0 für maximale Bandbreite und Zukunftssicherheit

Migration & technische Herausforderungen

- **Netzwerkdesign:**
  Vollständig dokumentierte VLANs, redundantes Routing, Traffic-Isolierung –
  entscheidend für spätere Audits
- **Auditfähigkeit:**
  Abgleich aller sicherheitsrelevanten Parameter (Firmware, BIOS, TPM, Signaturmodule) mit GSMA-Vorgaben
- **Performance:**
  Benchmarking der NVMe-Konfigurationen zur Sicherstellung stabiler eUICC-Signaturen unter hoher Last

Warum AMD EPYC?

- Maximale Rechenleistung: Bis zu 128 Kerne für HPC, Virtualisierung und Container.
- Modernste Schnittstellen: PCIe 5.0 und DDR5 für höchste Bandbreite.
- Skalierbar & zukunftssicher: Von Midrange- bis GPU-Cluster-Systemen.
- Planbare Kosten & volle Kontrolle: Souveräne On-Premises-Infrastruktur
  ohne Abhängigkeit von Hyperscalern.

Ergebnis & Lessons Learned

Ergebnis

Mit der neuen Infrastruktur verfügt der IT-Dienstleister über eine **skalierbare, hochverfügbare Cloud-Plattform**, die:

- Workloads flexibel verteilt und Latenzen minimiert,
- durch All-Flash-Storage konstante I/O-Performance garantiert,
- mit PCIe 5.0 und modularer Architektur optimal für zukünftige Erweiterungen vorbereitet ist,
- und gleichzeitig den Anforderungen von Datenschutz und Compliance gerecht wird.

Lessons Learned

Das Projekt unterstreicht, dass **On-Premises-Cloudplattformen auf AMD EPYC Basis** eine leistungsstarke Alternative zu Public-Cloud-Angeboten darstellen:

- **Datenhoheit und Sicherheit** bleiben unter voller Kontrolle des Betreibers.
- **Performance und Zuverlässigkeit** sind durch dedizierte Hardware jederzeit gewährleistet.
- **Skalierbarkeit** ermöglicht einen sukzessiven Ausbau, ohne bestehende Strukturen zu gefährden.

Jetzt beraten lassen

Rolle von Memorysolution

Memorysolution begleitete das Projekt von der Hardwareauswahl über die Feinkonzeption bis zur Bereitstellung und Vorkonfiguration der Systeme. Gemeinsam mit Mustang Systems wurde die Plattform rackfertig integriert und umfassend getestet, sodass der Kunde eine sofort einsatzbereite Lösung erhielt. Mit zusätzlichem Support- und Garantiepaket sicherte Memorysolution den stabilen Betrieb der Cloud-Infrastruktur langfristig ab.

Zur vollständigen Success Story

b) Projektbeispiel 2: Hochleistungsserver für KI und Visual Computing

Ausgangssituation

Ein international führendes Technologieunternehmen im Bereich optischer Erfassung und Verarbeitung visueller Informationen stand vor der Herausforderung, seine Entwicklungs- und Analyseumgebungen grundlegend zu modernisieren. Das Unternehmen entwickelt hochspezialisierte Kamerasysteme, Sensorik und Softwarelösungen, die Bild- und Sensordaten in Echtzeit auswerten – eingesetzt u. a. in industrieller Automatisierung, Forschung, Sicherheitstechnik und Medienproduktion.

Für neue Lösungen wie KI-gestützte Objekterkennung, 3D-Visualisierung oder hochauflösende Bildanalyse waren klassische Serverplattformen nicht mehr ausreichend: Es wurden enorme Rechenleistung, eine hohe Speicherbandbreite und eine schnelle, ausfallsichere Datenanbindung benötigt.

Umsetzung in der Praxis

Die Entscheidung fiel bewusst auf eine **On-Premises-Lösung**, um maximale Performance, Datenkontrolle und Verfügbarkeit zu gewährleisten. Zum Einsatz kam ein **Supermicro A+ GPU-Server 5126GS-TNRT2**, optimiert für KI- und Visual-Workloads.

Die Plattform wurde mit folgenden Kernkomponenten realisiert:

- **2× AMD EPYC 9355** (32 Kerne, 64 Threads, PCIe 5.0, 12-Kanal DDR5-6000) für höchste Parallelisierung und genügend PCIe-Lanes zur GPU-Anbindung.
- **10× NVIDIA RTX Pro 6000 Blackwell Server Edition** mit je 96 GB GDDR7 ECC für KI-Training, Rendering und Echtzeitanalyse.
- **1,5 TB DDR5-6400 ECC Registered DRAM** zur Unterstützung speicherintensiver Workloads.
- **NVMe-Storage mit 15,36 TB Samsung PM9A3 U.2 SSD** (mit Power-Loss-Protection und Hardwareverschlüsselung) sowie zusätzlicher Micron 7450 PRO für Systembetrieb.
- **Dual-100 GbE-Netzwerkschnittstellen** für die direkte Integration in hochperformante Daten- und Forschungsnetzwerke.
- Ergänzt durch ein umfassendes Servicepaket mit SLA (5/9/NBD), um Ausfallzeiten im kritischen Entwicklungsbetrieb zu vermeiden.

Die Lösung wurde vollständig rackoptimiert, redundant angebunden und bereits für den späteren **GPU-Ausbau** vorbereitet.

PCIe 5.0 & DDR5 – Schlüssel für KI-Workloads

Bei KI-Training und Visual Computing entscheidet die Bandbreite zwischen CPU, GPU und Speicher. PCIe 5.0 verdoppelt die Übertragungsrate gegenüber der Vorgängergeneration, während DDR5 mehr Kanäle und höhere Taktfrequenzen bietet. So lassen sich Datenströme effizient verarbeiten – mit dem Ergebnis: kürzere Trainingszeiten, geringere Latenzen und zukunftssichere On-Premises-Systeme.

Ergebnis & Lessons Learned

Ergebnis

Die neue Infrastruktur ermöglicht es, große Bild- und Sensordatenmengen in Echtzeit zu verarbeiten, komplexe KI-Modelle schneller zu trainieren und Visualisierungen mit höchster Detailtreue zu rendern. Damit konnte der Entwicklungsprozess deutlich beschleunigt und die Innovationskraft des Unternehmens nachhaltig gestärkt werden.

Lessons Learned

Das Projekt verdeutlicht, dass bei rechen- und datenintensiven Workloads lokale On-Premises-Infrastrukturen strategische Vorteile bieten:

- **Maximale Datenhoheit und Sicherheit,** da sensible Entwicklungsdaten nicht in externe Clouds ausgelagert werden müssen.
- **Planbare Kosten und garantierte Performance**, unabhängig von variablen Cloud-Kosten oder Provider-Limits.
- **Zukunftssichere Erweiterbarkeit,** da die Plattform modular skalierbar bleibt.

Die Kombination aus AMD EPYC Prozessoren, High-End-GPUs und NVMe-Speicher bildet die Grundlage für souveräne, leistungsfähige KI- und Visual-Computing-Umgebungen – direkt vor Ort, mit voller Kontrolle.

Rolle von Memorysolution

Jetzt beraten lassen

Memorysolution übernahm die technische Konzeption, die Auswahl der optimalen Server- und Speicherkomponenten sowie die Integration der GPU-Plattform. Gemeinsam mit Mustang Systems wurde die Lösung vorkonfiguriert, getestet und als schlüsselfertiges System ausgeliefert. Durch enge Abstimmung mit dem Kunden konnten die spezifischen Anforderungen an Datenanbindung, Performance und Verfügbarkeit exakt umgesetzt werden. Ergänzend stellte Memorysolution ein erweitertes Supportpaket mit SLA sicher, sodass der Kunde im laufenden Betrieb auf schnelle Reaktionszeiten und kompetente Unterstützung zurückgreifen kann.

Zur vollständigen Success Story

c) Projektbeispiel 3: GPU-beschleunigte HPC-Infrastruktur für Forschung & Bildgebung

Ausgangssituation

Eine führende Forschungseinrichtung mit Schwerpunkt auf Kommunikationstechnologien, Röntgentechnik und adaptiven Systemen stand vor der Herausforderung, ihre bestehende HPC-Umgebung deutlich zu erweitern. In Laboren und Rechenzentren entstehen enorme Datenmengen – von hochauflösender Röntgenbildgebung über komplexe Sensorsysteme bis hin zu Simulationen adaptiver Systeme. Parallel dazu wuchs die Bedeutung von KI-gestützten Methoden wie Deep Learning für Mustererkennung, Prognosen und Prozessoptimierung.

Gefordert war eine Lösung, die sowohl wissenschaftliche Simulationen als auch KI-Workloads souverän bewältigt – mit hoher Leistung, Stabilität und Zukunftssicherheit.

Umsetzung in der Praxis

Die Entscheidung fiel auf eine On-Premises-HPC-Lösung, die maximale Kontrolle über Daten und Prozesse bietet und zugleich modular erweiterbar bleibt. Realisiert wurde ein GPU-Cluster auf Basis des Supermicro A+ Server AS-5126GS-TNRT, ausgestattet für Hochlastbetrieb im 24/7-Modus.

Zentrale Komponenten::

- **2× AMD EPYC 9655** (96 Kerne, PCIe 5.0, 12-Kanal DDR5-6000) für maximale Parallelisierung und Bandbreite.
- **2,3 TB DDR5-6400 ECC Registered DRAM** für speicherintensive Simulationen und Datenpipelines.
- **8× NVIDIA Tesla H200 NVL** (141 GB HBM3e pro GPU, NVLink mit 900 GB/s) für GPU-zu-GPU-Kommunikation und verteiltes Training.
- **3× 15,36 TB Samsung PM9A3 U.2 NVMe SSDs** für schnellen, ausfallsicheren Datenzugriff.
- ISO9001/ESD-konforme Assemblierung und umfassende SLA-Absicherung für zuverlässigen Dauerbetrieb.

Warum On-Premises für Forschung unverzichtbar ist

In der Forschung entstehen hochsensible Daten, von medizinischer Bildgebung bis zu proprietären Algorithmen. Public-Cloud-Lösungen stoßen hier schnell an Grenzen. Eine On-Premises-HPC-Plattform gewährleistet volle Datenhoheit, garantiert deterministische Performance unabhängig von Provider-Auslastungen und ermöglicht klare Kostenkontrolle durch planbare Hardware-Investitionen. Das macht sie zur idealen Basis für souveräne wissenschaftliche Exzellenz.

Ergebnis & Lessons Learned

Ergebnisse:

Mit der neuen Plattform konnte die Einrichtung:

- Rechenzeiten für Simulationen und KI-Trainings drastisch verkürzen,
- große Datensätze in bislang unerreichter Geschwindigkeit verarbeiten,
- Forschungszyklen deutlich beschleunigen und neue Ansätze schneller in die Praxis überführen,
- durch PCIe 5.0 und modulare Architektur Zukunftssicherheit gewährleisten,

und einen stabilen 24/7-Betrieb sicherstellen.

Lessons Learned:

Das Projekt zeigt eindrücklich, dass GPU-optimierte On-Premises-Infrastrukturen für Forschung und Entwicklung strategische Vorteile bringen:

- Datenhoheit und Sicherheit bleiben vollständig gewahrt.
- Performance und Planbarkeit lassen sich unabhängig von Cloud-Kosten oder Provider-Limits sicherstellen.
- Zukunftssicherheit ist durch modulare Erweiterbarkeit und moderne Schnittstellen gegeben.

Die Kombination aus AMD EPYC Prozessoren, modernsten GPUs und NVMe-Speicher bietet die Grundlage für souveräne, leistungsfähige HPC-Umgebungen – maßgeschneidert für wissenschaftliche Exzellenz.

Rolle von Memorysolution

Jetzt beraten lassen

Memorysolution war von Anfang an in Beratung, Auslegung, Lieferung und Integration involviert. Die Systeme wurden im Rahmen des Mustang Systems Programms als betriebsbereite, dokumentierte Plattform geliefert – mit klarer Strategie für Performance, Sicherheit und Erweiterbarkeit.

So entstand eine souveräne Forschungsinfrastruktur, die High-End-GPU-Rechenleistung mit voller Kontrolle über Betriebs- und Datenprozesse vereint.

Handlungsempfehlungen & Checkliste

Abschließend fassen wir Empfehlungen zusammen, die IT-Entscheider:innen bei der Gestaltung ihrer Cloud-/On-Prem-Strategie berücksichtigen sollten. Diese Checkliste soll Ihnen helfen, nichts Wichtiges zu vergesse:

Strategische Planung

Workload-Analyse durchführen

- Inventarisieren Sie alle Anwendungen/Workloads und klassifizieren Sie sie nach Anforderungen (Datenschutz, Performance, Latenz, Verfügbarkeit).
- Fragen Sie: Welche müssen aus regulatorischen Gründen on-prem bleiben? Welche profitieren von Cloud-Diensten? Treffen Sie danach eine bewusste Platzierungsentscheidung pro Workload.

TCO-Kalkulation anstellen

- Berechnen Sie für wesentliche Systeme die 3-5-Jahres-Kosten in verschiedenen Szenarien (Cloud vs. On-Prem). Berücksichtigen Sie alle Kosten – inkl. Personal, Lizenz, Strom – und führen Sie Sensitivitätsanalysen durch (z.B. bei 50 % Auslastung vs. 100 %). Ermitteln Sie den Break-even-Punkt für Ihre typischen Lastprofile.
- Diese Zahlen helfen, fundiert argumentieren zu können.

Datenklassifizierung & Compliance-Map

- Etablieren Sie eine Datenklassifizierung (öffentlich, intern, vertraulich, kritisch…) und legen Sie fest, welche Klassen wo verarbeitet/gespeichert werden dürfen (Cloud öffentlich vs. Cloud mit Einschränkungen vs. nur On-Prem). Beziehen Sie Rechts- und Datenschutzabteilung hier ein. Diese Policies geben allen Beteiligten Klarheit (Entwicklung, Betrieb, Einkauf).

Partner/Anbieter sorgfältig wählen

- Suchen Sie sich vertrauenswürdige Partner für die Bereiche, die nicht Ihr Kerngeschäft sind. Für On-Prem-Hardware und Integration können Anbieter wie Memorysolution und Menzel IT Lösungen maßschneidern – nutzen Sie deren Expertise.
- Prüfen Sie Cloud-Anbieter auf regionale Angebote (gibt es z.B. einen Sovereign Cloud- Ableger oder lokale Rechenzentren). Achten Sie auf Zertifizierungen (BSI C5, ISO etc.).

Lock-in vermeiden

- Definieren Sie als Architekturprinzip, wann Sie auf Managed Services setzen und wann lieber auf offene Lösungen. Komplett auf proprietäre Platform Services zu setzen kann spätere Repatriierung erschweren.
- Nutzen Sie möglichst technologieagnostische Standards (z.B. Kubernetes, PostgreSQL, etc.), die Sie notfalls auch woanders betreiben können.
- Wenn Sie spezielle Cloud-Services einsetzen (weil sie einen großen Vorteil bringen), planen Sie zumindest grob einen Exit-Plan („was wäre nötig, um das on-prem/EU zu betreiben?“).

Umsetzungsphase

Pilotprojekte & schrittweises Vorgehen:
Führen Sie neue Ansätze erst in kleinerem Umfang ein (z. B. Non-Prod-App in Cloud, neue On-Prem-Cloud im Pilot). Schrittweises Deployment reduziert Risiken.

Netzwerk & Connectivity planen:
Bandbreite, Latenz, Redundanz frühzeitig adressieren. Dedizierte Cloud-Anbindung (ab mittlerem Datenvolumen meist sinnvoll). VPN/Direct Connect so konfigurieren, dass QoS und Monitoring möglich sind.

Security by Design:
Zentralisiertes Identity Management (Cloud Directory als Hub), MFA aktivieren, Daten verschlüsseln (in Transit und in Ruhe). Netzsegmentierung, Zero-Trust-Prinzip, Logging & Monitoring übergreifend implementieren.

Automatisierung & Tools:
Infrastructure-as-Code (z. B. Terraform) einsetzen, CI/CD-Pipelines vereinheitlichen (multi-cloud-fähige Tools). Automatisierung reduziert Fehler und schafft Reproduzierbarkeit.

Monitoring & FinOps implementieren:
Zentrales Monitoring vor Go-Live. Metriken für Performance und Kosten definieren. Logging-Ende-zu-Ende denken. FinOps-Prozess etablieren (monatliche Reports, Quartalsreviews).

Betrieb & kontinuierliche Verbesserung

Skill-Management:
Team-Skills regelmäßig prüfen und entwickeln. Trainings & Labs planen. Cross-Training zwischen Cloud- und On-Prem-Team fördern. MSPs als temporäre oder dauerhafte Unterstützung erwägen.

Regelmäßige Architektur-Reviews:
Strategie jährlich prüfen: Neue Services? Kostenveränderungen? Regulierung? Anforderungen? Workload-Verlagerung mitdenken.

Notfallplanung:
DR- und BC-Konzepte für beide Welten. Ausfallszenarien simulieren (z. B. Cloud-Ausfall, Offline-Modus). Szenarien regelmäßig testen.

Vermeidung von Schatten-IT:
Attraktive Self-Service-Angebote schaffen, Governance klären, Cloud-Abonnements zentral bekannt machen. Kultur fördern, in der Bedürfnisse offen kommuniziert werden.

Dokumentation & Wissenstransfer:
Aktuelle Architektur-, Datenfluss- und Verantwortlichkeitsdokumente pflegen. Wissenstransfer aktiv gestalten. Kollaborationstools für Wissensmanagement nutzen.

Diese Empfehlungen helfen, die vielen Aspekte geordnet anzugehen. Jedes Unternehmen hat andere Schwerpunkte – picken Sie die für Sie relevanten Punkte heraus und erweitern Sie die Checkliste ggf. um branchenspezifische Themen (etwa besondere Audit-Anforderungen).

Ausblick: Zukünftige Entwicklungen und Trends

Edge und 5G/6G Expansion

Mit 5G (und später 6G) steigt der Bedarf an Edge-Computing. Rechenleistung wandert näher an den Ort der Datenentstehung (autonome Fahrzeuge, Smart Cities, IoT). Erwartbar ist eine Verteilung hin zu dezentralen Knotenpunkten mit zentralem Management. Netzdesigns müssen das abbilden – viele verteilte Nodes, Orchestrierung über WAN, Zero-Trust-Security.

Souveräne KI & Datenräume

GAIA-X, europäische Datenräume und der AI Act verändern das Cloud-Nutzungsverhalten. Zertifizierte, souveräne Cloud-Angebote gewinnen an Relevanz. Kritische KI-Nutzung könnte künftig nur in regulierten Umgebungen erlaubt sein – Vorteil für On-Prem oder zertifizierte EU-Clouds.

Weiterentwicklung der Hardware

Quantencomputing bleibt mittelfristig Cloud-basiert, aber lokale Nutzung denkbar. Storage-Class-Memory verändert Datenverarbeitung. ARM-Server (z. B. Ampere, Nvidia Grace) ermöglichen dichte, stromsparende On-Prem-Systeme. Cloud-Kalkulationen müssen sich daran anpassen.

Software-Defined Everything & Multi-Cloud-Management

Abstraktionsschichten über Clouds hinweg werden besser: Tools wie Crossplane, Boundary oder VMware HCX ermöglichen Cloud-Brokering. K8s Federation, Cluster API – Workloads migrieren dorthin, wo Kosten/Performance am besten sind. Voraussetzung: Containerisierung & Standardisierung.

Security & Gesetzeslage

Geopolitik, Supply-Chain-Risiken, Gesetzesänderungen (z. B. strengere EU-Datenlokalisierung) machen On-Prem zum strategischen Rückzugsraum. Regelmäßige Audits und Zertifizierungen gewinnen an Bedeutung. Worst-Case-Szenarien (Cloud-Datenrückholung) durchdenken.

Kostenentwicklung & Marktbewegung

Cloud-Optimierung wird zum Dauerbrenner (FinOps). As-a-Service-Angebote wie HPE GreenLake oder Dell APEX etablieren sich on-prem. Das verschiebt die wirtschaftlichen Grenzen – RZ-Kapazität wie Cloud nutzbar, aber lokal kontrolliert.

Human Factor

Das Mindset wird pragmatischer. Cloud ist kein Allheilmittel, On-Prem kein Anachronismus. Der Bedarf steigt nach Generalist:innen, die beides können – Cloud-Architektur und Systembetrieb.

Fazit: Cloud vs. On-Prem bleibt ein Sowohl-als-Auch

IT-Architekturen der Zukunft sind hybrid, dynamisch und müssen souverän sein. Wer strategisch plant, flexibel bleibt und Partner wie Memorysolution und Menzel IT an seiner Seite weiß, kann heute schon die IT-Realität von morgen gestalten – resilient, effizient und nutzerzentriert.

Quellenverzeichnis

Hivenet: Understanding European tech sovereignty: why Europe is taking back control
https://www.hivenet.com/post/understanding-european-tech-sovereignty-why-europe-is-taking-back-control#:~:text=%2A%20Gaia,in%202022%29.%20%28European%20Parliament

Barclays CIO Survey, 1H 2024: 1H24 CIO Survey: 2024 Outlook Sustained
https://8198920.fs1.hubspotusercontent-na1.net/hubfs/8198920/Barclays_Cio_Survey_2024-1.pdf

Data Centre Review – „The next chapter of cloud in Europe“
https://datacentrereview.com/2025/05/the-next-chapter-of-cloud-in-europe/?utm_source=chatgpt.com

„FinOps in Focus 2025“ von Harness:
FinOps in Focus 2025 Control Cloud costs from code to Production
https://cdn.prod.website-files.com/6222ca42ea87e1bd1aa1d10c/67be20d4204f8f764a4410fa_FinOps%20in%20Focus%20Report.pdf

Private Cloud Outlook 2025 Report von Broadcom
https://www.vmware.com/docs/private-cloud-outlook-2025

Canalys Cloud Service Spending 2025:
Worldwide cloud service spending to grow by 19% in 2025
https://canalys.com/newsroom/worldwide-cloud-service-q4-2024

Flexera (State of the Cloud Report 2025)
https://info.flexera.com/CM-REPORT-State-of-the-Cloud-DE?lead_source=Organic%20Search

TCO-Analyse von Lenovo für KI-Hardware:
On-Premise vs Cloud: Generative AI Total Cost of Ownership Positioning Information
https://lenovopress.lenovo.com/lp2225-on-premise-vs-cloud-generative-ai-total-cost-of-ownership

McKinsey Europe Cloud Survey (2024):
The state of cloud computing in Europe: Increasing adoption, low returns, huge potential
https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-state-of-cloud-computing-in-europe-increasing-adoption-low-returns-huge-potential

Memorysolution GmbH
Hafenstr. 17, 79206 Breisach

Tel: +49 (0) 4298 95693 0
www.memorysolution.de
info@memorysolution.de