Bewertung des geplanten AZ-Redesign der PSP-Cloud

Stellungnahme

Version	00.04
Materialnummer	
Sprache	DE
Geschäftsbereich	GB 8

ROHDE&SCHWARZ

Inhaltsverzeichnis

Änderungshistorie

Datum	ÄI	Bemerkung	Geändert durch
jjjj-mm-tt	nn.nn		Abt/Name
2021-10-15	00.01	Initiale Version erstellt und übergeben	8IPC-SA/Lars Wächtler
2021-10-29	00.02	Review und Überarbeitung	8IPPP/Daniel Breest
2021-11-08	00.03	div. Überarbeitungen, insb. Kapitel 7.8.3 sowie 7.11.4	8IPC-SA/Lars Wächtler
2021-11-10	00.04	Fehlerbehebung bei internen Referenzen	8IPC-SA/Lars Wächtler

Prüfverzeichnis

Die folgende Tabelle zeigt einen Überblick über alle Prüfungen - sowohl Eigenprüfungen als auch Prüfungen durch die interne Qualitätssicherung - des vorliegenden Dokuments:

Datum	Geprüfte ÄI	Bemerkung	Prüfer
jjjj-mm-tt	nn.nn		Abt/Name
2021-10-29	00.02	Eigenprüfung	8IPPP/Daniel Breest
2021-11-08	00.03	Eigenprüfung	8IPPP/Daniel Breest
2021-11-10	00.04	Eigenprüfung	8IPC-SA/Lars Wächtler

Glossar

Ein Glossar mit Begriffen, die häufig verwendet werden, befindet sich im R&S®COMVIDENCE Glossar. Die dortigen Definitionen gelten auch in diesem Dokument.

1 Einleitung/Auftrag

Die Abteilung IT des BKAs beabsichtigt, ein Re-Design der Verfügbarkeitszonen in der BKA-Cloud vorzunehmen: Von – Stand heute – zwei Rechenzentren mit je drei Verfügbarkeitszonen auf drei Rechenzentren mit je einer Verfügbarkeitszone. Im Rahmen des Änderungsverfahrens CR-43/2021 wurde Rohde & Schwarz damit beauftragt das neue Design hinsichtlich der Auswirkungen auf die Projektlaufzeit, die Projektkosten und Projektrisiken vor bzw. nach der Wirkbetriebsaufnahme von PHOENIX mit Release 1.1 zu bewerten.

Zur Vorbereitung fand ein dreitägiger Workshop vom 05.10.2021 bis zum 07.10.2021 am BKA-Standort Wiesbaden mit den relevanten Serviceprovidern des BKA statt, insbesondere mit dem PSP-Team als Betreiber der BKA-Cloud und Initiator des Re-Designs der Verfügbarkeitszonen. Die wesentlichen Ziele des Workshops waren, das neue Design detailliert zu erörtern und mögliche Problemstellungen zu identifizieren. Das Protokoll dieses Workshops ist im Projekt-Confluence des BKAs unter https://confluence.psp.bka.bund.de/x/NYXLAg bzw. https://confluence.psp.bka.bund.de/x/G4DLAg abgelegt und dient unter anderem diesem Positionspapier als Grundlage und Referenz.

Basierend auf den innerhalb des Workshops gewonnenen Erkenntnissen hat Rohde & Schwarz anschließend die Bewertung des neuen Designs vorgenommen, welche in diesem Dokument dargelegt und diskutiert werden.

An dieser Stelle muss allerdings angemerkt werden, dass einige Aspekte, wie z.B. die Verfügbarkeit des dritten, externen Rechenzentrums sowie des Geo-Redundanten Standortes B* aktuell noch nicht final geklärt ist und das Rohe & Schwarz an dieser Stelle nur mögliche Optionen skizzieren kann deren konkrete Bewertung bzw. Aussagen erst nach Vorliegen aller notwendigen Informationen möglich ist.

Im Folgenden werden in den Kapiteln 4 und 5 zur Heranleitung der Bewertung die beiden Designs, sowohl das aktuelle Design mit drei Regionen, wovon zwei Regionen auf die Brandabschnitte I und II am Standort Wiesbaden abgebildet sind, als auch das neue Design der BKA Cloud mit drei Rechenzentren als Verfügbarkeitszonen in der Region Rhein Main technisch skizziert.

Basierend auf dieser Diskussion wird im Kapitel 6 detailliert auf verschiedene Aspekte des neuen Designs eingegangen, um diese im Kapitel 2 final zu bewerten.

2 Zusammenfassung

Nach Einschätzung von Rohde & Schwarz lassen sich nach aktuellem Kenntnisstand keine signifikanten Auswirkungen auf die Projektlaufzeit erkennen, wenn die Umsetzung des neuen Designs vor Beginn der Abnahmetests von Release 1.1 abgeschlossen ist. Sollte die Umsetzung des neuen Designs in das Zeitfenster der Abnahmetests fallen, sind entsprechend der Dauer der Umbauten bzw. Migration bezüglich dieser Tests Verzögerungen zu erwarten, was aus Sicht von Rohde & Schwarz sehr wahrscheinlich zu einem Verzug der Wirkbetriebsaufnahme führen und einem Projektverzug führen würde.

Wenn die Umsetzung des neuen Designs nach Wirkbetriebsaufnahme erfolgt, ist aus Sicht von Rohde & Schwarz zwar kein Projektverzug erkennbar. Allerdings erhöht sich dadurch das Projektrisiko z.B. durch Ausfall des bzw. Datenverlust im Produktivsystems. Deshalb sollte von dieser Option Abstand genommen werden.

Im Fall der Umsetzung des neuen Designs erwartet Rohde & Schwarz, bis auf den oben diskutierten möglichen Verzug sowie der in Kapitel 7.8 diskutierten Konsistenzproblematik des S3-Objektspeichers kein erhöhtes Projektrisiko durch das neue Design. Maßgeblich für diese Einschätzungen ist die Zusicherung der verschiedenen Serviceprovider des BKAs für die Fachanwendung PHOENIX, dass alle bisher geltenden Servicegarantien auch mit dem neuen Design eingehalten werden.

Da sich das neue Design an Realisierungs- bzw. Industriestandards gängiger öffentlicher und privater Infrastruktur- bzw. Cloudanbieter orientiert, wird dieses Design sowohl durch die, von der Fachanwendung benötigten Basistechnologien[1], der PaaS-Umgebung (TKGm) als auch der Fachanwendung selbst unterstützt. Maßgeblich für diese Einschätzung sin die im Workshop gegebenen Servicegarantieren bzgl. der zugesicherten maximalen Latenz und Bandbreite zwischen den einzelnen Komponenten innerhalb einer Region und der Vollvermaschung der Regionen bzw. Rechenzentren.

Zur Sicherstellung der Verfügbarkeit der Fachanwendung PHOENIX im Falle eines Ereignisses, welches zum Komplettausfall des Standortes Wiesbaden führt, ist ein zweiter georedundanter Standort notwendig. Diese Feststellung gilt aber praktisch sowohl für das aktuelle als auch für das neue Design und ist für die Gesamtbewertung neutral.

Aufgrund der aktuellen technischen Nähe der beiden Regionen Wiesbaden I und Wiesbaden II im aktuellen Design werden die technischen und fachlichen Querschnittsdienste bzw. externen Dienste[2] wie in Kapitel 7.4 dargestellt, nicht für jede Region unabhängig bereitgestellt. Basierend auf der Realisierung dieser Dienste sowie der zugesicherten Basisservicegarantieren sind nach aktueller Einschätzung von Rohde & Schwarz keine Auswirkungen erkennbar. Im Gegenteil, das neue Design behebt diese Schwachstelle des aktuellen Designs und führt aus Sicht von Rohde & Schwarz zu einem konsistenteren Gesamtsystem.

Beiden Designs, ist gemein, dass die Fachanwendung PHOENIX von den konkreten physikalischen Gegebenheiten mithilfe des SDDC-Konzepts abstrahiert wird und die gleichen Servicegarantien zugesichert werden. Ausgehend davon sind durch die Umstellung auf das neue Design keine Auswirkungen auf zentrale Konzepte[3] für das Fachverfahren zu erwarten.

Weiterhin geht Rohde & Schwarz davon aus, dass die Umstellung auf das neue Design keine signifikanten Änderungen an der Software R&S COMVIDENCE notwendig macht, da es aus Sicht der Anwendung keine sichtbaren Unterschiede unter Berücksichtigung der aktuell zugesicherten Servicegarantien gibt. Aufgrund der Reduktion der zu unterstützenden Regionen, erwartet Rohde & Schwarz dadurch tendenziell eher eine Reduktion der Komplexität insbesondere in den Bereichen Geo-Replikation, Systemadministration und Deployment, was die Robustheit des Fachverfahrens PHOENIX insgesamt erhöht.

Für die PHOENIX-Cloud sind Anpassungen insofern notwendig, dass die bestehende Basis-Netzwerkinfrastruktur auf das neue Design umgestellt und erweitert werden muß. In diesem Zusammenhang erfolgt auch die Umstellung der von Rohde & Schwarz im Rahmen der Betriebsunterstützung bereitgestellten und betriebenen Komponenten auf das neue Design. Steht die Basisinfrastruktur der PSP bereit, geht Rohde & Schwarz davon aus, dass alle Services, welche in der Verantwortung von Rohde & Schwarz liegen, innerhalb weniger Wochen entsprechend neu bereitgestellt werden können. Um an dieser Stelle keine sichtbaren Auswirkungen auf die Tests der gelieferten Inkremente zu generieren, geht Rohde & Schwarz davon aus, dass es entweder zeitlich begrenzt möglich ist Infrastrukturressourcen der PSP für die entsprechende Migration einzusetzen oder dass die Migration so eingeplant wird, dass die Migration und der damit verbundene Verlust der PHOENIX-Cloud in ein Zeitfenster fällt, in dem keine Aktivitäten auf der PHOENIX-Cloud vorgesehen bzw. notwendig sind.

Das neue Design wirkt sich positiv auf die Projektkosten aus, dass es am Standort Wiesbaden zu einer Halbierung der notwendigen Datenreplikate von sechs im aktuellen Design auf drei im neuen Design kommt. Dies reduziert den Hardware-Ressourcenbedarf entsprechend. Im Bereich der Applikationsinstanzen ist von keiner signifikanten Reduktion auszugehen, da deren Anzahl eher von der Arbeitslast als von der Anzahl der Regionen abhängt. Weiterhin reduzieren sich voraussichtlich die operativen Kosten für den Betrieb der Fachanwendung am Standort Wiesbaden, da nur noch eine Region betrieben werden muss.

Auf der Gegenseite stehen die notwendigen Kosten zur Anpassung des Pflichtenhefts, zum Umbau der Betriebsumgebung sowie die Kosten für die potentielle Beschaffung einer weiteren Dell EMC ECS-Instanz für die dritte Verfügbarkeitszone anfallen. Weitere Kosten, z.B. für weitere Lizenzen, sowohl der PHOENIX Cloud als auch bei der PHOENIX Fachanwendung sind aktuell nicht bekannt. Im Rahmen der zweiten Phase, wie in

[1] Apache Kafka, Apache Cassandra, Elastic Search und weitere
[2] z.B. DNS, GIS oder Verzeichnisdienste,
[3] u.a. IT-Sicherheitskonzept, Netzwerk- und Netzwerksegmentierungskonzept

den Kapiteln 5 Vorstellung des geplanten Designs der PSP-Cloud auf Seite 9 und 6 Zeitliche Betrachtungen auf Seite 2 diskutiert, des neuen Designs können zur Erhöhung der Verfügbarkeit zusätzliche Kosten entstehen, z.B. durch Bereitstellung eines weiteren Zugangspunktes zu den Verpflichteten für das externe dritte Rechenzentrum der Region Rhein-Main. Da diese aber aktuell nicht konkret geplant ist, ist eine Bewertung zum jetzigen Zeitpunkt durch Rohde & Schwarz nicht realisierbar und wird deshalb nur als potentielle Kosten dargestellt.

3 Hauptrisiken

- Unklare Rechtslage bzgl. des Betriebs der externen Verfügbarkeitszone durch R&S,
- Notfallstromversorgung von W2 und W2-H2O
- Datenkonsistenz des S3-Speichers

4 Vorstellung des aktuellen Designs der PSP-Cloud

An das Fachverfahren PHOENIX werden unter anderem hohe bis sehr hohe Anforderungen im Bereich der Verfügbarkeit sowohl innerhalb einer Region als auch über mehrere Regionen verteilt gestellt. Aufgrund dieser Verfügbarkeitsanforderungen wurden für die Laufzeit und zum Persistieren von Daten Technologien vorgegeben, wie z.B. Pivotal Cloud Foundry bzw. TKGm, bzw. selektiert, wie z.B. Apache Kafka oder Apache Cassandra, welche zur Erreichung der Verfügbarkeit mehrere Instanzen einer Applikation bzw. mehrere Replikate der Daten in mehreren unabhängigen Umgebungen, den Verfügbarkeitszonen, voraussetzen. Diese verteilt arbeitenden Systeme benötigen intern zur Sicherstellung der Verfügbarkeit und der Konsistenz der Daten eine ungerade Anzahl von Instanzen bzw. Replikaten, wobei z.B. bei drei Instanzen bzw. Replikaten der Verlust von einer Instanz bzw. eines Replikats kompensiert werden kann.

Aufgrund der Gegebenheiten mit nur zwei Rechenzentren am Standort des BKAs konnten diese notwendigen Minimalanforderungen nicht über Rechenzentren, wie sonst allgemein üblich, realisiert werden. Aus diesem Grund wurden die z.B. die Pivotal Cloud Foundry Foundations jeweils in die beiden Rechenzentren am Standort des BKAs deployed, so dass diese aus Sicht der Fachanwendung PHOENIX als zwei unabhängige Regionen zu betrachten sind. Die weiterhin notwendigen Verfügbarkeitszonen wurden auf die vSAN Fault Domains abgebildet, wie in den Kapiteln 5 Hochverfügbarkeit und Georedundanz und 6 Umsetzung der BKA-Cloud des BKA Cloud Grobkonzepts Version 1.0 vom 23.03.2018 diskutiert. Diese Aufteilung wurde Rohde & Schwarz als betriebliche Vorgabe zur Realisierung der Applikation gemacht.

Basierend auf dieser Vorgabe und der Notwendigkeit von mindestens drei Verfügbarkeitszonen für die verschiedenen Datenhaltungskomponenten des Fachverfahrens sowie der Forderung nach einem realen georedundanten Standort wurde das Fachverfahren PHOENIX so konzipiert, dass jedes Rechenzentrum am Standort Wiesbaden als eine eigene, unabhängige Region angenommen wurde, um sowohl die Betriebs- als auch die Entwicklungskomplexität beherrschbar zu halten. Die Abbildung 1 stellt diesen Zusammenhang schematisch dar.

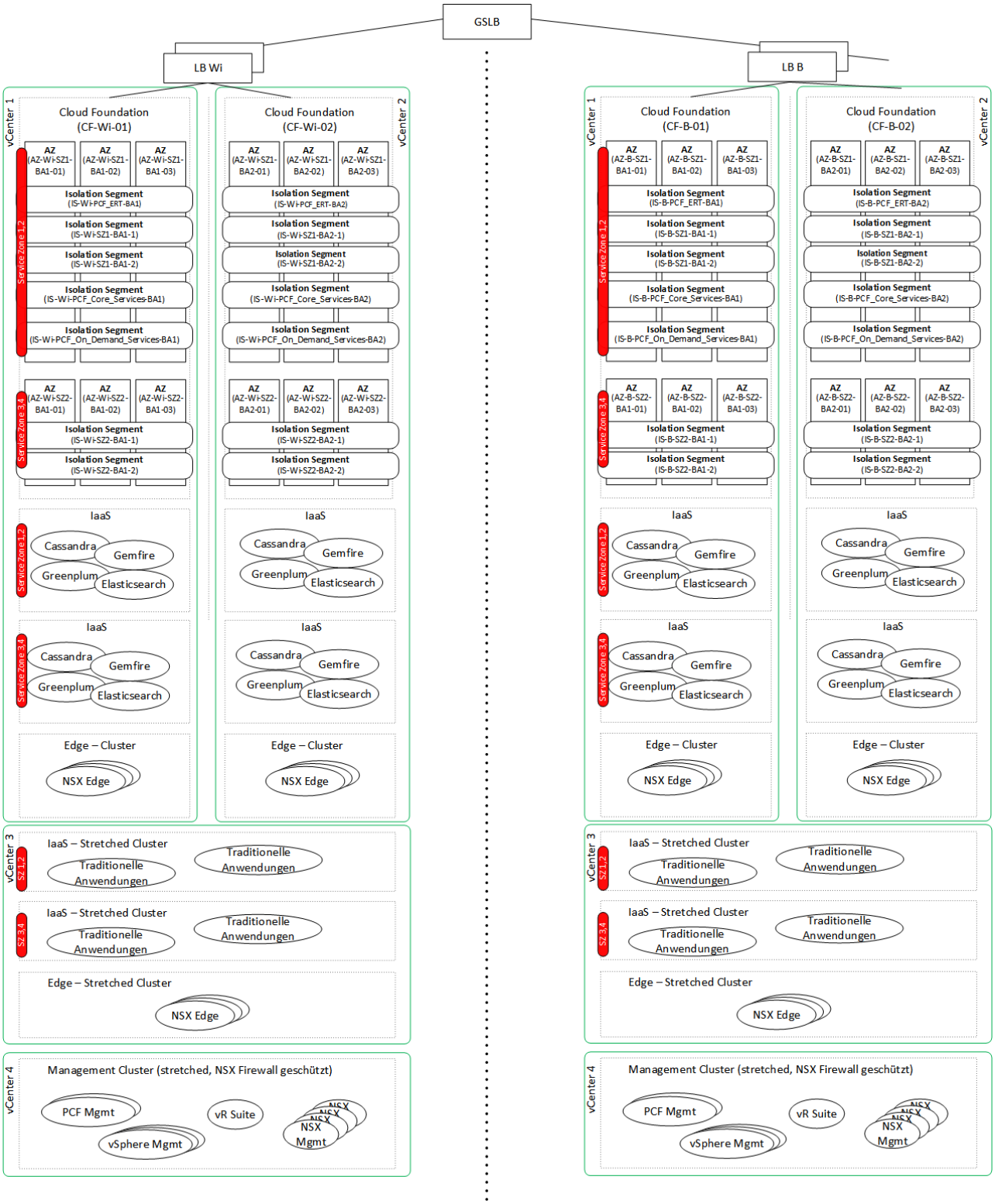

Abbildung 1: Cluster Design BKA Cloud (in Anlehnung an: Abbildung 6-1 aus dem BKA Cloud Grobkonzept)

Basierend darauf wird jedes Rechenzentrum für PHOENIX zunächst als eigenständig betrachtet, wobei die Daten zwischen den Rechenzentren über einen zusätzlichen Mechanismus repliziert werden. (Pflichtenheft PHOENIX v8.30, REQ-26037). Eine Ausnahme zu dieser Art der Replikation stellen administrative Daten dar, welche innerhalb eines relationalen Datenbankmanagement Systems (engl. Relational Database Management System, RDBMS) persistiert und repliziert werden.

Ausgehend von diesen Vereinbarungen aus dem Pflichtenheft erwartet das Fachverfahren PHOENIX alle bereitgestellten Querschnittsdienste, in denen Daten persistiert werden, unabhängig in jedem Rechenzentrum.

5 Vorstellung des geplanten Designs der PSP-Cloud

Das geplante Design erweitert die aktuell zwei Rechenzentren Wiesbaden I und Wiesbaden II um mindestens ein weiteres Rechenzentrum, so dass in der Region Rhein-Main dann mindestens drei Rechenzentren und somit drei Verfügbarkeitszonen realisiert werden können.

In der ersten Realisierungsphase wird ein weiteres virtuelles Rechenzentrum, Wiesbaden II-H2O, bereitgestellt. Dieses Rechenzentrum ist physikalisch im Rechenzentrum Wiesbaden II verortet, ist aber, bis auf wenige Ausnahmen, wie z.B. den Notfalldieselgenerator, autark.

In einer weiteren Realisierungsphase soll dann ein weiteres Rechenzentrum in synchroner Kommunikationsreichweite, d.h., mit einer garantierten maximalen Netzwerklatenz von weniger als 5 ms, bei einem externen Dienstleister entstehen.

Aus rein funktioneller Sicht der Fachanwendung innerhalb einer Region, im aktuellen Design innerhalb eines Rechenzentrums, verändert sich mit dem neuen Design bis auf eine geänderte Servicegarantie des Dell EMC ECS Objektspeichers im Bereich der Konsistenz nichts, da die Fachanwendung direkt von den drei vSAN Fault Domains innerhalb eines Rechenzentrums auf die drei Rechenzentren abgebildet werden kann.

Wenn die nichtfunktionalen Anforderungen der Fachanwendung in den Dimensionen Verfügbarkeit, Vertraulichkeit, Integrität und Performance mit dem neuen Design ebenfalls sowohl von der Betriebsplattform als auch von den Querschnittsdiensten erfüllt werden, ist die Abbildung der Fachanwendung PHOENIX auf das neue Design in diesem Bereich ebenfalls transparent.

Die Abbildung 2 stellt das alte Design und das neue Design aus Sicht der Fachanwendung PHOENIX schematisch einander gegenüber.

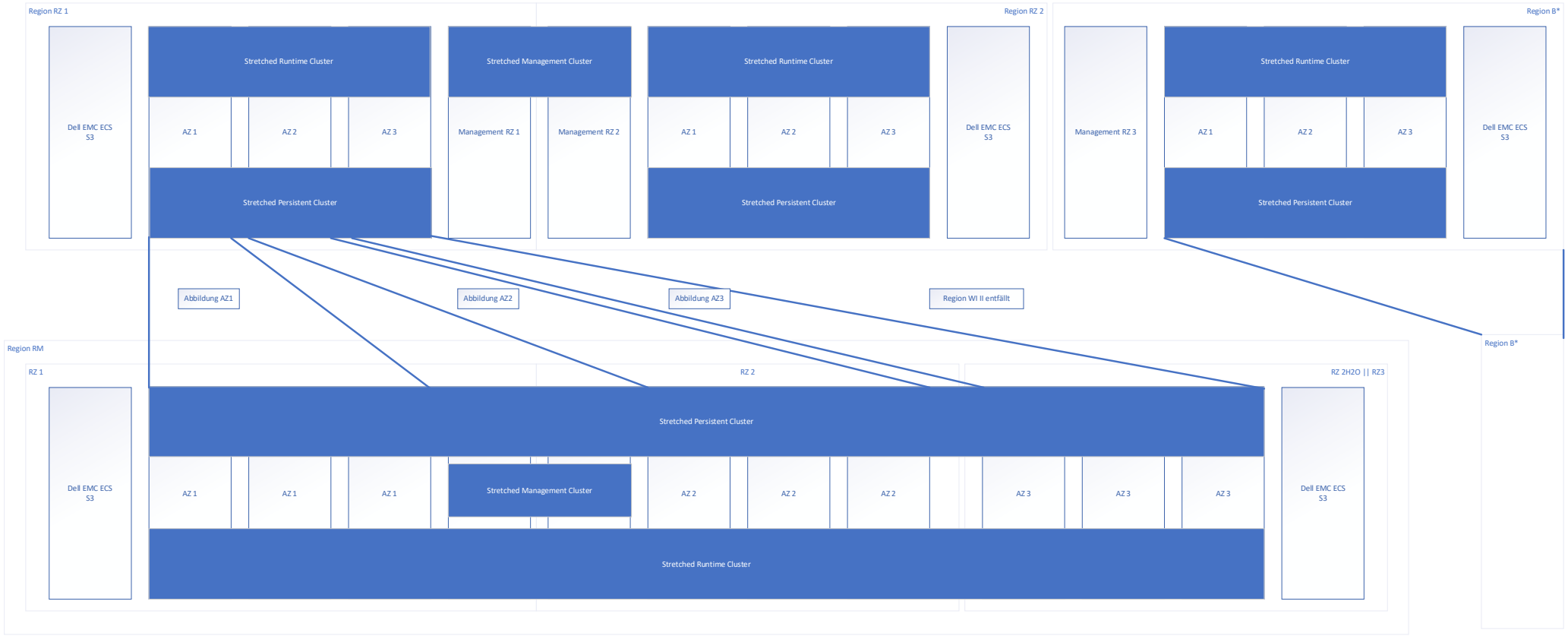

Abbildung 2: Vergleich des aktuellen und des neuen Designs der Verfügbarkeitszonen

Im Zentrum der Darstellung wird ersichtlich, dass die Region Wiesbaden II des aktuellen Designs ersatzlos entfällt. Die Verführbarkeitszonen und somit deren Services der Region Wiesbaden I werden auf die Rechenzentren der Region Rhein-Main abgebildet. Die Geo-Region B* ist in beiden Designs äquivalent und ihr kommt in beiden Designs die gleiche Wichtigkeit zu, da dies die einzig verbliebene weitere Region darstellt.

Man muss an dieser Stelle konstatieren, dass die Regionen Wiesbaden I und Wiesbaden II aus technischen Notwendigkeiten entstanden sind und somit keinen reellen Schutz gegen Katastrophen am Standort Wiesbaden bieten. Aus diesem Gesichtspunkt ändert sich die Notwendigkeit einer Geo-Region zwar theoretisch, aber nicht praktisch.

Die Abbildung 3 illustriert das neue Design aus Sicht eines *Apache Cassandra Clusters* und einer Applikation innerhalb eines *Workload Clusters*. Dabei ist es unerheblich, ob es sich um eine Pivotal Cloud Foundry oder um ein TKGm Kubernetes Workload Cluster handelt.

Relevant ist, dass die einzelnen Daten *Data A* und *Data B* innerhalb des *Apache Cassandra Clusters* dreimal repliziert vorhanden und dass die einzelnen Replikate auf die verschiedenen Verfügbarkeitszonen, in der Abbildung als *Availability Zone* bezeichnet, verteilt sind. Innerhalb einer Verfügbarkeitszone werden die Daten weiterhin auf die zu dieser Verfügbarkeitszone gehörenden Server verteilt.

Äquivalent zu den Daten des *Apache Cassandra Clusters* ist eine Applikation, in der Abbildung als *Client* bezeichnet, im *Workload Cluster* dargestellt, welche ebenfalls, um die geforderte Verfügbarkeit zu gewährleisten, mehrfach instanziiert ist, wobei die einzelnen Instanzen ebenfalls auf die Verfügbarkeitszonen aufgeteilt sind.

Darüber hinaus ist die physikalische Netzwerkinfrastruktur dargestellt. Die *Leaf Switches* stellen innerhalb der Fault Domain eine Layer 2 basierte Paketvermittlung (L2) bereit. Diese *Leaf Switches* sind dann mehrfach redundant an die in den einzelnen Rechenzentren verorteten *Spine Switches* angebunden und folgen somit dem klassischen Spine-Leaf Ansatz. Die *Spine Switches* der Rechenzentren sind dann ebenfalls mehrfach redundant an die *Spine Switches* der *Core Network Infrastructure* angebunden. Dieses Netzwerkdesign stellt sicher, dass die einzelnen Services der verschiedenen Rechenzentren, wie z.B. der *Client Instance III* oder die Nodes des *Apache Cassandra Clusters*, unabhängige Netzwerkpfade zwischen den Verfügbarkeitszonen besitzen.

Die *Spine Switches* stellen dabei, unabhängig von der logischen Verortung in einer Verfügbarkeitszone bzw. der *Core Network Infrastructure* eine Layer 3 basierte Paketvermittlung (L3) bereit.

Sollte es zu einem Verlust einer Verfügbarkeitszone, in der Abbildung 3 die *Availability Zone II*, kommen, können aufgrund der Datenredundanz, für das Fachverfahren PHOENIX wird im Allgemeinen eine Redundanz innerhalb einer Region von drei angenommen, innerhalb des bzw. der Cluster die Anfragen der *Clients* weiterhin beantwortet werden. Die dazu notwendige interne Kommunikation des *Apache Cassandra* Clusters wird, unabhängig vom Ausfall aller Services inklusive der Netzwerkinfrastruktur der *Availability Zone II*, über die als logisch unabhängig zu betrachtende *Core Network Infrastructure* realisiert.

Unter der Voraussetzung, dass die zugesicherten Servicegarantien im Bereich der Betriebsplattform und den Querschnittsdiensten des aktuellen Designs auch für das neue Design gelten, ist das neue Design innerhalb einer Region funktional äquivalent zum aktuellen Design.

ROHDE&SCHWARZ

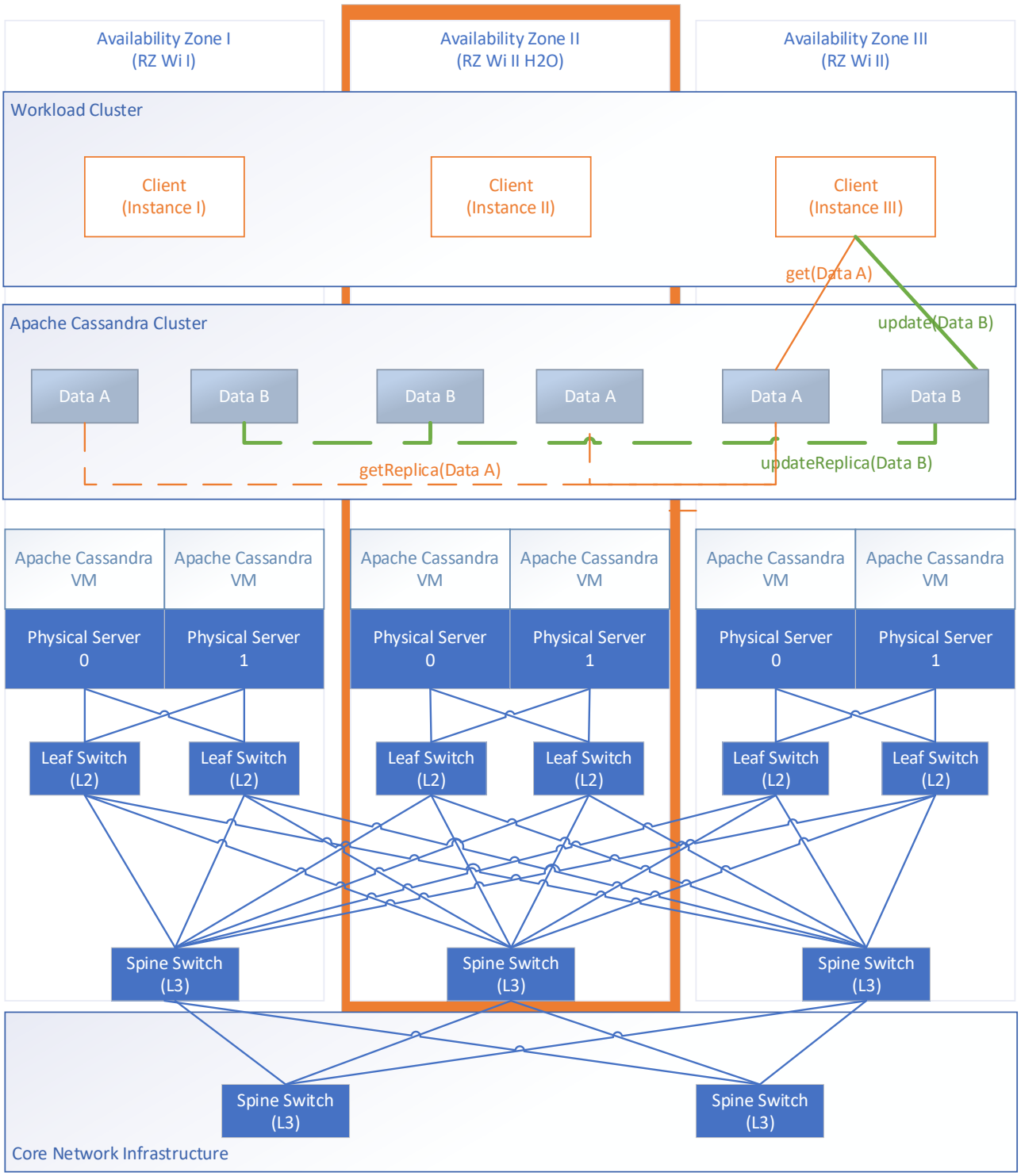

Abbildung 3: Neues Design am Beispiel des Apache Cassandra Clusters und einer Applikation im Workload Cluster

Wichtig ist dabei, dass Rohde & Schwarz, im Gegensatz zum Vorschlag der PSP davon ausgeht, dass die Datenhaltung weiterhin als Services auf virtuellen Maschinen und nicht als Workloads innerhalb eines Kubernetes-Clusters realisiert wird. Eine Diskussion der möglichen Konsequenzen der Umstellung des Betriebsmodells persistierender Service hin zu einem Kubernetes-basierten Modells ist nicht Teil dieser Betrachtungen.

ROHDE&SCHWARZ

6 Zeitliche Betrachtungen

Aus Sicht von R&S sollte die Wirkbetriebsaufnahme Ende September 2022[4] bereits auf einer Cloud-Plattform nach dem neuen Verfügbarkeitszonen-Konzept (AZ-Redesign) erfolgen. Hintergrund ist, dass nach dem aktuellen Konzept separate Instanzen von R&S COMVIDENCE in den beiden Rechenzentren WI-1 und WI-2 aufgesetzt werden müssten, welche dann zu einem späteren Zeitpunkt auf das neue Konzept migriert werden müssten.

Da die Realisierung der benötigten dritten Verfügbarkeitszone in einem externen Rechenzentrum in synchroner Entfernung frühestens im dritten Quartal 2022, eventuell auch erst im ersten Quartal 2023 erfolgen wird[5], und damit erst nach Bereitstellung von Release 1.1, wird der Aufbau einer 3. Verfügbarkeitszone in WI-H2O zur vorübergehenden Nutzung angestrebt. Die Umsetzung dieses Schrittes erfolgt frühestens Mitte November 2021 nach der Stellungnahme von Rohde & Schwarz und der PG-PHOENIX. Die Umbauarbeiten werden von der PSP-Cloud verantwortet und voraussichtlich bis Ende des ersten Quartals 2022 abgeschlossen sein. Eine feste Zusage eines Fertigstellungsdatums durch die PSP-Cloud steht allerdings noch aus.

Daraus ergibt sich die Realisierung des AZ-Redesigns in zwei Phasen:

1) Um- und Ausbau der Region Rhein-Main unterteilt in
    a. Aufbau der 3. Verfügbarkeitszone in WI-H2O zur vorübergehenden Nutzung und mit reduzierter Ausfallsicherheit bzw. Verfügbarkeit,
    b. Aufbau der 3. Verfügbarkeitszone in einem externen Rechenzentrum in synchroner Reichweite zu den beiden Verfügbarkeitszonen der Region Rhein-Main in Wiesbaden (Zielausbau),
2) Aufbau der 2. geo-redundanten Region am Standort B*.

Abbildung 4 stellt die wichtigen Meilensteine des AZ-Redesigns ins Verhältnis zu den definierten Releases von R&S COMVIDENCE.

Tabelle 1 stellt die Meilensteine mit Beschreibung und Datum dar. Letztere basieren auf den im Workshop kommunizierten bzw. in Aussicht gestellten Terminen.

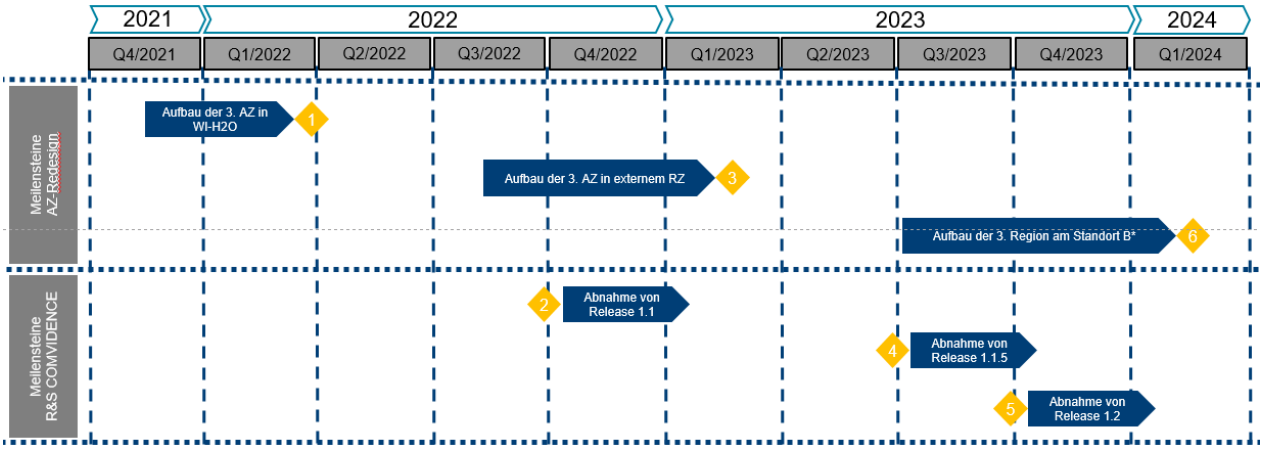

Abbildung 4: Meilensteine des AZ-Redesigns und Meilensteine der R&S COMVIDENCE-Releases

[4] Unter Wirkbetriebsaufnahme wird hier die Bereitstellung des Release 1.1 und der Beginn der Abnahme verstanden.
[5] Aussage der Projektgruppe Geo im Workshop vom 5. – 7. Oktober 2021 in Wiesbaden

ROHDE&SCHWARZ

Meilenstein	Beschreibung	Datum
1	Der Aufbau der 3. Verfügbarkeitszone in W2-H2O zur vorübergehenden Nutzung ist abgeschlossen.	Ende Q1/2022
2	Bereitstellung von Release 1.1 von R&S COMVIDENCE	30.09.2022
3	Der Aufbau der 3. Verfügbarkeitszone in einem externen Rechenzentrum in synchroner Reichweite ist abgeschlossen.	Mitte Q1/2023
4	Bereitstellung von Release 1.1.5 von R&S COMVIDENCE	30.06.2023
5	Bereitstellung von Release 1.2 von R&S COMVIDENCE	30.09.2023
6	Der Aufbau der 2. Region am Standort B* ist abgeschlossen.	Mitte Q1/2024

Tabelle 1: Übersicht der Meilensteine

6.1.1 Um- und Ausbau der Region Rhein-Main

Die Bereitstellung einer dritten Verfügbarkeitszone in einem externen Rechenzentrum in synchroner Reichweite vor Wirkbetriebsaufnahme von PHOENIX ist der präferierte Weg, da dabei aus Sicht des Projektes PHOENIX keine zusätzlichen Aufwände beispielsweise für Migration anfallen. Allerdings ist diese Variante auf Grund der zeitlichen Gegebenheiten – Aussage der Projektgruppe Geo bzgl. Verfügbarkeit der 3. AZ - aus Sicht von Rohde & Schwarz nicht realistisch.

Die Realisierung des neuen Designs am Standort Wiesbaden soll deshalb in zwei Schritten erfolgen. Im ersten Schritt wird die notwendige dritte Verfügbarkeitszone im wassergekühlten Teil des Brandabschnittes 2 am Standort Wiesbaden zur vorübergehenden Nutzung durch PHOENIX realisiert (W2-H2O). Im zweiten Schritt erfolgt dann im Rahmen des Multi Cloud-Projekts der Umzug in das Rechenzentrum eines externen Dienstleisters. Dieses Rechenzentrum muss sich in synchroner Reichweite befinden und Netzwerklatenzen von weniger als 5 ms garantieren.[6]

Im Rahmen dieser Auslagerung der dritten Verfügbarkeitszone müssen die Daten bzw. Instanzen aus dem wassergekühlten Teil des Brandabschnittes 2 am Standort Wiesbaden an den externen Standort verlagert werden. Da aktuell die konkreten Gegebenheiten noch nicht bekannt sind, sollen im Weiteren die beiden wahrscheinlichsten Optionen dafür diskutiert werden:

1) die logische Übertragung bzw. der physikalische Transport der Daten und Instanzen an den neuen Standort oder
2) der vorübergehende Betrieb von vier Verfügbarkeitszonen.

Beide Optionen besitzen individuelle Vor- und Nachteile.

Bei der ersten Option wird für den Zeitraum der logischen Übertragung bzw. des physikalischen Transports sowie die notwendigen Vor- und Nacharbeiten die Verfügbarkeit der Fachanwendung insoweit reduziert, dass der Ausfall einer weiteren Komponente nicht mehr kompensiert werden kann und die Fachanwendung komplett ausfällt. Auf der anderen Seite benötigt diese Option nur minimale zusätzliche Hardware-Ressourcen.

Beim vorübergehenden Betrieb einer vierten Verfügbarkeitszonen wird die Verfügbarkeitszone in W2-H2O temporär auf den vierten, externen Standort erweitert und die einzelnen Workload-Instanzen dieser neuen Verfügbarkeitszone entsprechend migriert. Dazu werden entsprechend Ressourcen für den Betrieb dieser vierten Verfügbarkeitszone, z.B. im Rechenzentrum des externen Dienstleisters, sowie die notwendige Netzwerkanbindung bzw. Netzwerkbandbreite benötigt. Bei dieser Option ist die Verfügbarkeit der Fachanwendung während der Migration nicht beeinträchtigt.

[6] VMware vSAN-Anforderung

ROHDE&SCHWARZ

6.1.2 Aufbau der zweiten georedundanten Region

Die Realisierung der zweiten georedundanten Region wird laut Projektgruppe Geo erst in 2023/24 erfolgen. Die fehlende Georedundanz stellt ein hohes Risiko für die Verfügbarkeit der Fachanwendung dar. Dieses Risiko besteht aber unabhängig vom AZ-Design. Das Risiko könnte mit zwei weiteren Verfügbarkeitszonen in synchroner Distanz minimiert werden. Dies würde jedoch weitere Kosten für das Projekt bedeuten.

7 Technische Bewertung des neuen Designs

Das neue Design der Verfügbarkeitszonen kann von der Fachanwendung PHOENIX unter Berücksichtigung verschiedener funktionaler und nichtfunktionaler Servicegarantien der Betriebsplattform bzw. der Querschnittsdienste direkt verwendet werden.

Die zentrale funktionale Servicegarantie, damit die Fachanwendung das neue Design der Verfügbarkeitszonen in der Region Rhein-Main direkt nutzen kann, ist die Vollvermaschung der Rechenzentren bzw. Verfügbarkeitszonen untereinander, so dass der Ausfall eines Rechenzentrums innerhalb einer Region nicht zum Verlust weiterer Verfügbarkeitszonen und somit zum Verlust der gesamten Region führt. Eine weitere zentrale funktionale Servicegarantie der Betriebsplattform ist die Erreichbarkeit der notwendigen fachlichen und technischen Querschnittsdienste.

Die zentral notwendige nichtfunktionale Servicegarantie stellt in diesem Kontext Anforderungen an die Latenz sowie Bandbreite für die Netzwerkkommunikation zwischen den verschiedenen Rechenzentren.

Darüber hinaus stellt die Verfügbarkeit der von der Fachanwendung PHOENIX benötigten Services der Betriebsplattform bzw. der Querschnittsdienste eine weitere nichtfunktionale Anforderung dar. Diese Anforderung bestand allerdings bereits im aktuellen Design und muss im Rahmen der Realisierung des neuen Designs entsprechend wieder mit realisiert werden.

7.1 Auswirkungen auf die notwendige Hardware und Software

Im Rahmen des geplanten Re-Designs sollen die drei Verfügbarkeitszonen aus der bestehenden Hardware der beiden Rechenzentren WI-1 und WI-2 sowie der bereits für das Auftragnehmer-Referenzsystem beschafften Hardware aufgebaut werden. Dies hat Einfluss auf die verwendete Hardware und Software-Lizenzen. Im Folgenden wird der Einfluss auf die benötigte Hard- und Software sowie etwaige Kosten oder Einsparungen diskutiert.

7.1.1 Hardware

Im Rahmen der Erstellung der dritten Verfügbarkeitszone wird sowohl Server- als auch weitere Unterstützungshardware, wie z.B. Switches benötigt. Um die Kosten an dieser Stelle so neutral wie möglich zu halten, ist die Verwendung der bereits für das Auftragnehmer-Referenzsystem beschafften Hardware von der PSP vorgesehen, da diese exakt die notwendigen Ressourcen bietet um die dritte Verfügbarkeitszone infrastrukturell auszustatten. Des Weiteren können die aus dem Auftragnehmer-Referenzsystem freiwerdenden Edge- und Management-Server als Erweiterung für den Edge- bzw. Management-Cluster der neuen Verfügbarkeitszone verwendet werden.

Für die Redundanz des Objektspeichers der dritten Verfügbarkeitszone ist es in Phase 2 der Realisierung des neuen Designs vorgesehen, eine dritte Dell EMC ECS Instanz zu beschaffen. Hierdurch werden entsprechend Kosten für diesen Objektspeicher entstehen. Da es aufgrund der in Kapitel 7.7 dargestellten Probleme hinsichtlich Stromversorgung, Kühlung sowie Platz am Standort Wiesbaden nicht möglich ist,

ROHDE&SCHWARZ

diese dritte Dell EMC ECS Instanz am Standort Wiesbaden zu betreiben, ist aber abzusehen, dass die dritte Instanz des Objektspeichers frühestens im Rahmen der zweiten Phase des neuen Designs beschafft werden wird.

7.1.2 Software und Lizenzen

Äquivalent zur Wiederverwendung der Hardware des Auftragnehmer-Referenzsystems ist die entsprechende Wiederverwendung der bereits dafür beschafften Softwarelizenzen geplant.

Es ist allerdings aktuell nicht geklärt, ob bereits entsprechende TKGm-Lizenzen für das Auftragnehmer-Referenzsystem beschafft wurden. Die Beschaffung dieser Lizenzen ist jedoch unabhängig von dieser Diskussion notwendig und soll somit hier nicht weiter betrachtet.

Daher werden nach aktuellem Kenntnisstand von Rohde & Schwarz keine weiteren Softwarelizenzen benötigt.

7.2 Notfallstromversorgung der Verfügbarkeitszonen W2 und W2-H2O

Die Verfügbarkeitszonen W2 und W2-H2O werden über denselben Dieselgenerator mit Notfallstrom versorgt. Dies kann zu einer reduzierten Verfügbarkeit der Fachanwendung führen, da von einem möglichen Ausfall des Dieselgenerators zwei Verfügbarkeitszonen betroffen sind. Dieses Problem wird erst mit der Bereitstellung der dritten Verfügbarkeitszone in einem entfernten Rechenzentrum in synchroner Distanz behoben. Die reduzierte Verfügbarkeit liegt nach aktuellem Stand ab Aufnahme des Wirkbetriebs mit Release 1.1 am 30.09.2022 bis zur externen Bereitstellung der 3. Verfügbarkeitszone Mitte Q1/2023 vor (4,5 Monate).

Berücksichtig man, dass mehrere Fehlerereignisse – Stromausfall in Kombination mit Ausfall des Dieselgenerators - oder ein katastrophaler Fehler eintreten müssen, um das beschriebene Problem zu verursachen, wird das Risiko als gering eingeschätzt.

7.3 Anforderungen an die physikalische und virtuelle Netzwerkinfrastruktur

Das neue Design hat primär Auswirkungen auf die zugrundeliegenden physische als auch virtuelle Netzwerkinfrastruktur. Nachfolgend wird auf diese Punkte eingegangen.

7.3.1 Änderungen des physikalischen Netzwerk-Designs

Das im Rahmen des neuen Designs angedachte physikalische Netzwerkdesign erweitert die klassische Spine-Leaf-Architektur um eine weitere Spine-Ebene. Diese verbindet die einzelnen Rechenzentren der Region Rhein-Main miteinander. Dabei stellen die Leaf-Switches innerhalb der Fault Domain eine Layer 2-basierte Paketvermittlung bereit. Diese Leaf-Switches sind mehrfach redundant an die Spine-Switches der einzelnen Rechenzentren angebunden und folgen dem klassischen Spine-Leaf-Ansatz. Die Spine-Switches der Rechenzentren sind dann ebenfalls mehrfach redundant an die Spine-Switches der Kern- Netzwerk-Infrastruktur angebunden. Dieses Netzwerkdesign stellt sicher, dass die einzelnen Services der verschiedenen Rechenzentren unabhängige Netzwerkpfade zwischen den Verfügbarkeitszonen bzw. den Rechenzentren besitzen.

Allen Spine-Switches ist gemein, dass sie unabhängig von der logischen Verortung in einer Verfügbarkeitszone bzw. der Kern-Netzwerk-Infrastruktur, eine Layer 3-basierte Paketvermittlung

ROHDE&SCHWARZ

bereitstellen. Sollte es zu einem Verlust einer Verfügbarkeitszone bzw. eines Rechenzentrums kommen, können aufgrund der entsprechenden Pfadredundanz die verbleibenden Rechenzentren weiterhin miteinander kommunizieren. Somit sind die Verfügbarkeitszonen untereinander voll vermascht.

Die Bereitstellung externen Services, wie z.B. den technischen als auch fachlichen Querschnittsdiensten erfolgt somit transparent für die Fachanwendung für alle Verfügbarkeitszonen.

7.3.2 Kapazität der physikalischen Netzwerk-Infrastruktur

Die physikalische Netzwerk-Infrastruktur wird vom Infrastruktur-Provider PSP-Cloud als leistungsfähig genug angesehen, um die Anwendungslasten der verschiedenen Fachanwendungen zu tragen, wobei sowohl im aktuellen als auch im geplanten neuen Design keine expliziten Quality of Service-Garantien gegeben werden.

Wenn die Kapazität der physikalischen Netzwerk-Infrastruktur ausgeschöpft ist, kann nach Aussage des Infrastruktur-Providers die Kapazität durch Schaltung weiterer physikalischer Verbindungen einfach erhöht werden (z.B. 2* 100GbE pro Link).

Dies wird auch für das in Kapitel TBD beschriebene externe Rechenzentrum in synchroner Distanz zugesichert. Hier ist vorgesehen, entsprechende Dark Fiber-Links in der benötigten Kapazität anzumieten und diese Anmietung bei Bedarf zu erweitern.

7.3.3 Latenzen

Die maximale Latenz innerhalb des SDDCs muss kleiner als 5ms sein, da bei einer höheren Latenz Probleme innerhalb der vSAN-Cluster auftreten. Es wird unterschieden in Ausbreitungslatenz, Übertragungslatenz und Latenz innerhalb von Netzwerkgeräten, wie Routern und Switches. Im Folgenden werden die Auswirkungen auf die drei Latenzarten diskutiert.

7.3.3.1 Ausbreitungslatenz

In ersten Schritt der Realisierung des neuen Designs – vorübergehender Aufbau der 3. AZ in W2-H2O - wird es keine Erhöhung der Ausbreitungslatenz und Übertragungslatenz geben, da der Aufbau lokal am Standort Wiesbaden mit derselben Leistung erfolgt.

Im zweiten Schritt wird die 3. Verfügbarkeitszone in einem externen Rechenzentrum in synchroner Distanz aufgebaut. Dies führt zu einer Erhöhung der Ausbreitungslatenz im Netzwerk, da die Entfernung bei rund 10km liegen wird. Diese Erhöhung ist zwar berechenbar, sollte aber in der Praxis keine messbaren Auswirkungen haben. Bei einer angenommenen Entfernung l von 10km und einer Lichtgeschwindigkeit c von 299.792.458m/s beträgt die Ausbreitungslatenz $t_{Ausbreitung}$

$$t_{Ausbreitung} = \frac{l}{2/3 * c}$$

$$t_{Ausbreitung} = \frac{3 * 10 * 10^3 ms}{2 * 299.792.458\text{m}}$$

$$t_{Ausbreitung} = 0{,}00005003461\text{s}$$

rechnerisch 0,0500346 ms.

Die Übertragungslatenz kann sich durch die Anmietung eines Rechenzentrums in synchroner Distanz erhöhen. Dies kann vermieden werden, wenn die Übertragungsgeschwindigkeit vergleichbar zu der lokalen Übertragungsgeschwindigkeit angemietet wird.

Die Latenz die durch Router, Switches oder andere Netzwerkgeräte verursacht wird, kann laut Aussage des Infrastruktur-Providers PSP-Cloud vernachlässigt werden und wird sich nicht auf die Gesamtlatenz auswirken.

Insgesamt kann somit eine Gesamtlatenz von unter 5 ms vom Infrastruktur-Serviceprovider als Voraussetzung für einen Rechenzentrums-übergreifenden Betrieb der vSAN-Cluster gewährleistet werden.

Für die Fachanwendung PHOENIX ist eine geringe Netzwerklatenz ebenfalls essentiell um die geforderten Servicegarantien, speziell im Bereich Performance zu gewährleisten. Nach aktueller Einschätzung von Rohde & Schwarz ändern sich somit die zugesicherten Servicegarantien im Bereich Latenz nicht für die Fachanwendung PHOENIX.

7.4 Querschnittsdienste und externe Dienste

Um den hohen bzw. sehr hohen Anforderung an die Verfügbarkeit der Fachanwendung gerecht zu werden und einen georedundanten Betrieb in mehreren Rechenzentren bzw. Regionen zu ermöglichen, ist die Fachanwendung PHOENIX bisher als eine Anzahl von Installationen konzipiert, welche in den einzelnen Rechenzentren autark betrieben werden (siehe REQ-34152 PHOENIX-Pflichtenhefts).

Um die Unabhängigkeit der einzelnen Installationen des Fachverfahrens in den einzelnen Rechenzentren bzw. Regionen zu gewährleisten, geht das Fachverfahren PHOENIX davon aus, dass auch die benötigten Querschnittsdienste autark je Rechenzentrum bzw. Region bereitgestellt werden.

Im Rahmen des Workshops zum Re-Design der Verfügbarkeitszonen vom 05.10.2021 bis zum 07.10.2021 hat sich allerdings herausgestellt, dass einige technische und fachliche Querschnittsdienste für die beiden nach altem Design geplanten Installationen des Fachverfahrens PHOENIX am Standort Wiesbaden nicht autark bereitgestellt werden bzw. bereitgestellt werden sollen. Speziell für solche Querschnittsdienste, bei denen Daten durch die Fachanwendung modifiziert werden, können daraus Inkonsistenzen der Daten führen, etwa dann wenn ein Objekt innerhalb eines solchen Querschnittsdienste bereits von einem PHOENIX-Service in einem anderen Rechenzentrum bzw. einer anderen Region erzeugt bzw. modifziert wurde.

Die skizzierte Problematik würde sich durch das neue Designs zumindest für die Region Rhein-Main dahingehend auflösen, dass es nur noch eine Installation der Fachanwendung gibt. Dies ist aus Sicht von Rohde & Schwarz ein klarer Vorteil des neuen Designs.

7.5 Diskussion Stretched VMware vSAN/vSphere Cluster

Im aktuellen Design war ein Stretched VMware vSAN-Cluster pro lokaler Region in der Planung vorgesehen. Die Legacy-Applikationen hätten so redundant pro lokale Region bereitgestellt werden sollen. Zum jetzigen Stand wurde die Redundanz der Stretch Cluster nicht umgesetzt. Mit dem neuen Design bleibt dieser Zustand bestehen und es wird nur ein Stretched vSAN Cluster für alle drei Verfügbarkeitszonen bereitgestellt. Die Fachanwendung selbst nutzt keinen Stretched vSAN Cluster direkt, sondern nur per Querschnittsdienst oder Endpunkt eines Legacy-Dienstes.

Dadurch dass es nur einen Stretched vSAN Cluster für alle Verfügbarkeitszonen der Region Rhein-Main geben soll, entsteht das Risiko einer Minimierung der Verfügbarkeit sollte der Stretched vSAN Clusters ausfallen. Dieses Risiko kann durch weitere Stretched vSAN Cluster verringert werden, wodurch Kosten für das Projekt entstehen. Das Risiko eines solchen Komplettausfalles ist als gering einzuschätzen und wird durch den geplanten georedundanten Standort weiter minimiert.

ROHDE&SCHWARZ

7.6 Auswirkungen des Re-Designs auf bestehende SLAs

An das Fachanwendung PHOENIX bzw. an dessen Teilverfahren werden fachlich folgende Anforderungen im Bereich der Verfügbarkeit gestellt. Diese sollen an dieser Stelle zur Herleitung der Verfügbarkeiten der technischen und fachlichen Querschnittsdienste dienen. Weitere Anforderungen, z.B. in der Dimension Performance, werden im Rahmen dieses Dokuments nicht betrachtet, da das Re-Design keine Auswirkungen auf diesen Faktor hat.

Das Fachverfahren PHOENIX zerfällt im Rahmen dieser Betrachtung in Teilverfahren 1 - Entgegennahme und temporäre Ablage der TKÜ-Rohdaten mit sehr hoher Verfügbarkeit und die Teilverfahren 2 bis 11 mit jeweils hoher Verfügbarkeit. Alle Komponenten des Teilverfahrens 1 müssen selbst mindestens sehr hoch verfügbar sein, um die Funktionen des Teilverfahrens in der geforderten Verfügbarkeit bereitstellen zu können. Exemplarisch sei an dieser Stelle die Laufzeitumgebung für das Teilverfahren, die Schnittstelle zu den Verpflichteten und die Kafka-Cluster für die Bereitstellung der Kafka-1-Warteschlange genannt.

Im Rahmen des Workshops zum Re-Design der Verfügbarkeitszonen wurde dem Projekt PHOENIX und Rohde & Schwarz als Lieferant der Fachanwendung PHOENIX von den beteiligten Serviceprovidern des BKAs versichert, dass die ursprünglich zugesagten Servicegarantien insbesondere im Hinblick auf die Verfügbarkeit auch mit dem neuen Design gewährleistet sind.

Eine Abweichung von bestehenden Servicegarantien ergibt sind nach Einschätzung von Rohde & Schwarz jedoch im Bereich der Datenkonsistenz des S3-Objektspeichers Dell EMC ECS. Im aktuellen Design ist eine strikte Konsistenz für den Objektspeicher gegeben. Es wird vermutet das dies im neuen Design nicht mehr der Fall ist. Der Serviceprovider des Objektspeichers prüft, welche Optionen zur Wiederherstellung der aktuellen Zusagen auch im neuen Design möglich sind.

Sollte es nicht möglich sein, die ursprünglichen Zusagen in Bezug auf die Datenkonsistenz des Objektspeicher innerhalb einer Region im neuen Design einzuhalten, besteht aus Sicht von Rohde & Schwarz ein hohes Risiko einer Inkonsistenz der Daten innerhalb einer Region. Dies ist besonders kritisch angesichts der im Objektspeicher abgelegten Datenarten TKÜ-Rohdaten und den daraus abgeleiteten Produktdaten. Eine Gewährleistung der Konsistenz könnte dann nur durch sehr komplexe Anpassungen auf Ebene der Fachanwendung kompensiert werden. Zudem ließen sich die geforderten harten Zeitlimits z.B. im Bereich der Live-Auswertung unter Umständen nicht mehr gewährleisten, weil die verarbeitenden Prozesse der Fachanwendung nicht sofort die gespeicherten Daten aus dem Objektspeicher lesen können (Read after Write).

7.7 Auswirkung auf die Klima- sowie Stromproblematik

Die Mehrheit der Daten des Fachverfahrens PHOENIX wird in Technologien wie Apache Cassandra oder Elasticsearch gespeichert, welche einen unterbrechungsfreien lesenden und schreibenden Zugriff erlauben. Dazu wird das eigentliche Datum mehrfach abgelegt. Im konkreten Design werden im Allgemeinen drei Replikate eines Datums innerhalb einer Region gesichert, was im aktuellen Design einem Rechenzentrum entspricht.

Da aufgrund der technischen Gegebenheiten, wie in Kapitel 4 diskutiert, im aktuellen Design jedes Rechenzentrum als eine eigene Region betrachtet wird, werden die Daten entsprechend in jeder Region mehrfach und unabhängig von anderen Regionen persistiert. Daraus folgt für das aktuelle Design, das die Daten am Standort Wiesbaden typischerweise sechsfach repliziert vorliegen, was einen entsprechenden Ressourcenbedarf mit sich bringt.[7]

Aufgrund der Reduktion der Regionen und damit einhergehend der Verfügbarkeitszonen im neuen Design reduziert sich somit auch die Anzahl der Datenreplikate um die Hälfte, was wiederum zu einer Reduktion

[7] Sechsfach, da im aktuellen Design die beiden RZs am Standort Wiesbaden als eigenständige Region mit je 3 Verfügbarkeitszonen betrachtet werden.

der benötigten Hardware-Ressourcen beiträgt bzw. die aktuell vorgesehenen Hardware-Ressourcen entsprechend länger den Ressourcenbedarf der Fachanwendung PHOENIX erfüllen können.

Dies wiederum entspannt das Problem der notwendigen Ressourcenbereitstellung für die Infrastruktur selbst in den Dimensionen

- verfügbare Höheneinheiten in den Brandabschnitten des Standortes Wiesbaden,
- notwendige Bereitstellung von Strom sowie
- notwendige Bereitstellung von Kühlung.

Die zwangsläufig notwendige Fixinfrastruktur, wie z.B. die Kern-Netzwerk-Infrastruktur, ist dazu im Verhältnis als gering anzusehen, was zu einer Verbesserung dieser Situation beiträgt.

7.8 Datenkonsistenz des S3-Objektspeichers

7.8.1 Bewertung des aktuellen Designs

Im aktuellen sowie im neuen Design herrscht eine klare Trennung der Regionen. Das heißt, alle Regionen und die darin laufenden Installationen/Instanzen der Fachanwendung PHOENIX (in Abbildung 1 violett markiert) sind eigenständig und können voneinander unabhängig betrachtet werden. Dies betrifft auch die Objektspeicher *Dell EMC ECS* (in Abbildung 1 rot markiert) der einzelnen Regionen, welche keine Daten synchronisieren und auch nicht aufeinander zugreifen bzw. sich kennen. Basierend darauf, dass innerhalb einer Region immer nur ein Objektspeicher existiert, ist im aktuellen Design die Konsistenz der innerhalb des Objektspeichers abgelegten Daten implizit gewährleistet. Schreibende Zugriffe, wie beispielsweise von der *Applikation A* (in Abbildung 1 grün markiert), finden auf der gleichen *Dell EMC ECS Instanz* statt wie lesende Zugriffe von *Applikation B* (in Abbildung 1 gelb markiert). Somit kann es per Definition nicht vorkommen, dass ein geschriebenes Objekt in einem anderen Zustand gelesen werden kann.

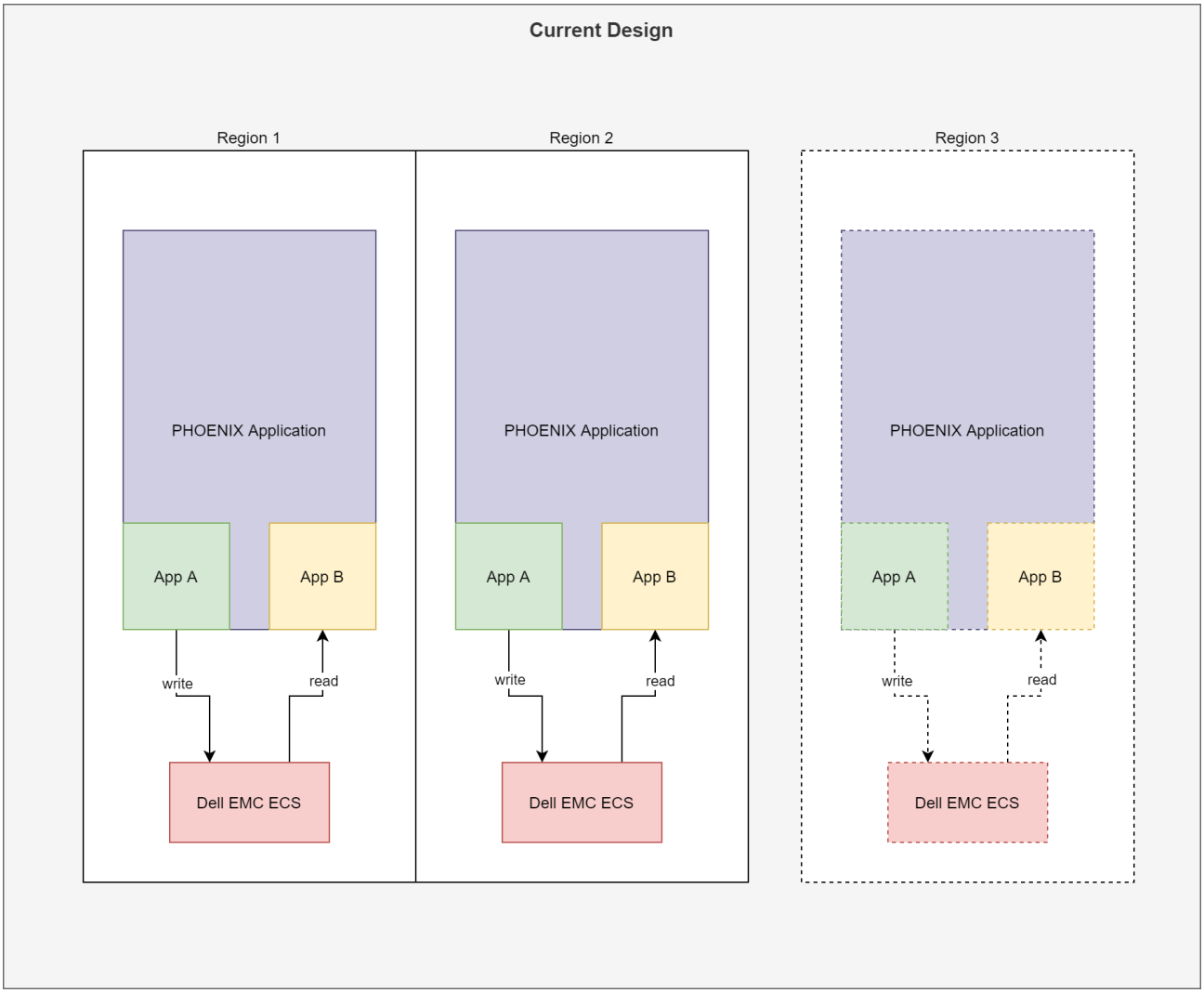

Abbildung 5: Dell EMC ECS im aktuellen Design (geplante Elemente sind mit gestrichelten Linien versehen)

7.8.2 Bewertung des neuen Designs

Das neue Design sieht eine vergleichbare Trennung für die Verfügbarkeitszonen am Standort Wiesbaden bisher nicht vor. Aufgrund der notwendigen Servicegarantien, speziell im Bereich Verfügbarkeit, müssen die Dell EMC ECS-Instanzen miteinander gekoppelt werden. Dabei wird technologiebedingt eine asynchrone Replikation der Objekte verwendet. Das bedeutet, das Objekte, die in eine Instanz geschrieben wurden, noch nicht auf einer anderen Instanz verfügbar bzw. in einer älteren Version verfügbar sein können. Diese Objekte werden erst dann verfügbar, wenn die asynchrone Replikation erfolgt ist. Durch dieses Replikations-Lag[8] entsteht eine Inkonsistenz zwischen den gekoppelten Instanzen. Infolgedessen wechselt der Objektspeicher von einem inhärent konsistenten Betriebsmodell im aktuellen Design zu einem eventuell konsistenten Betriebsmodell im neuen Design. Das neue Betriebsmodell ist ein vom Serviceproviders des Objektspeichers vorgeschlagenes Design.

Dies ist in der folgenden Abbildung 6 skizziert. Dabei schreibt die *Applikation A* (in Abbildung 6 grün dargestellt) ein Objekt in die erste *Dell EMC ECS Instanz* (in Abbildung 6 rot markiert) in der Verfügbarkeitszone 1 der Region 1. Wenn der Schreibvorgang erfolgreich abgeschlossen ist, folgt die asynchrone Replikation. Liest jedoch *Applikation B* (in Abbildung 6 gelb dargestellt) aus der Verfügbarkeitszone 2 vor Vollendung dieser Replikation von einer anderen *Dell EMC ECS Instanz,*

[8] Das Replikations-Lag beschreibt die Zeit, welche für die Replikation benötigt wird

bekommt diese Applikation B einen anderen Zustand des Objektes als Ergebnis. Dies passiert obwohl Applikation A und B in derselben Region und derselben *PHOENIX Applikation* laufen. Der Grund hierfür ist die neuartige Kopplung der *Dell EMC ECS Instanzen*, welche im aktuellen Design nicht existiert.

Diese Inkonsistenz kann zu problematischen Zuständen innerhalb der *Fachanwendung PHOENIX* führen.

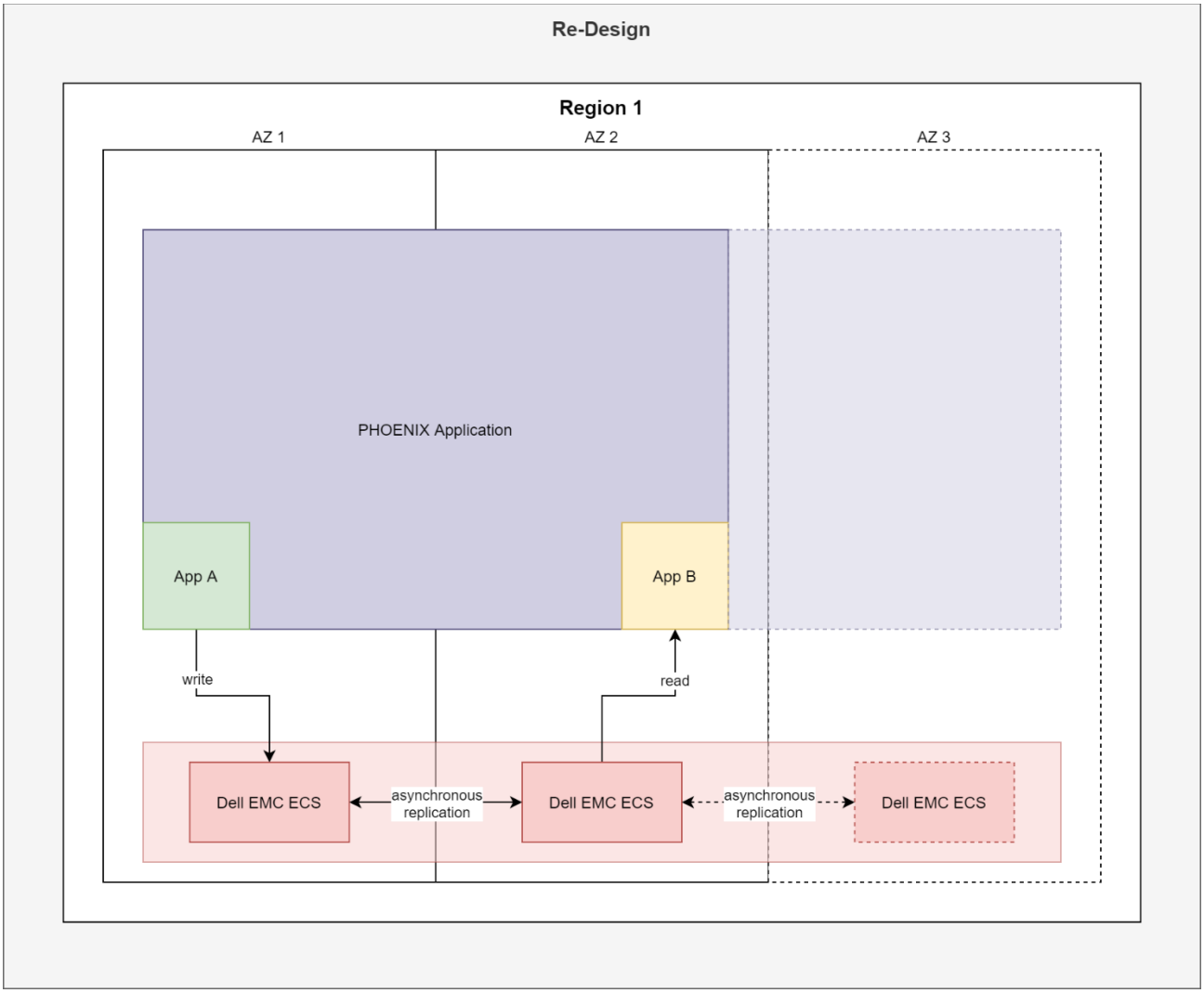

Abbildung 6: Dell EMC ECS im neuen Design (geplante Elemente sind mit gestrichelten Linien versehen)

Die somit notwendige Konsistenzprüfung der Objekte sollte aus Sicht des Serviceproviders vom Objektspeicher applikatorisch erfolgen. Hierzu muss die Applikation den Replikationsstatus eines Objektes bei jeden Schreib- bzw. Lesesvorgang prüfen, was zu signifikanten Mehraufwänden sowohl im Bereich der Entwicklung als auch der notwendigen physikalischen Ressourcen des *Fachverfahrens PHOENIX* führt. Zudem entsteht im neuen Design bzw. Betriebsmodell ein Replikations-Lag, welcher sich auf die Verarbeitungsleistung des *Fachverfahren PHOENIX* negativ auswirken wird. Diese Replikationsverzögerung ist im aktuellen Design innerhalb einer Region nicht vorhanden.

Um die aktuell zugesicherten Verfügbarkeiten weiterhin garantieren zu können, muss im neuen Design ein temporärer Teilausfall des Objektspeichers betrachtet werden. Im aktuellen Design würde die Region in der eine *Dell EMC ECS Instanz* ausfällt, komplett ausfallen wobei allerdings die zweite Region am Standort Wiesbaden diesen Ausfall entsprechend kompensieren kann. Dies ist nur möglich, da die beiden Regionen komplett autark konzipiert und die Objektspeicher nicht gekoppelt sind.

Das neue Design würde so einem multi-site ECS Deployment entsprechen. Hier kann ein sogenanntes Access During Outage (ADO) Feature genutzt werden. Dabei kann auf Daten zugegriffen werden, wenn ein Standort (in diesem Fall eine *Dell EMC ECS Instanz*) ausfallen würde (Temporary Site Outage). Das ADO

ROHDE&SCHWARZ

Feature wird auf Bucketebene konfiguriert und gewährt Zugriff auf Buckets, selbst wenn eine *Dell EMC ECS Instanz* nicht mehr erreichbar ist.

Somit könnte auch während des temporären Ausfalls einer ECS-Instanz weiterhin auf die Buckets in der anderen ECS-Instanz zugegriffen werden. Dabei ist zu beachten, dass bei Netzwerkausfall diese Lesezugriffe keine Änderungen auf Seiten des Originalbuckets beinhalten, die während des Ausfalls auftreten.

7.8.3 Vorschlag eines alternativen Designs

Um den oben genannten Bedenken entgegenzuwirken, schlägt Rohde & Schwarz folgendes Design in Abbildung 7 vor.

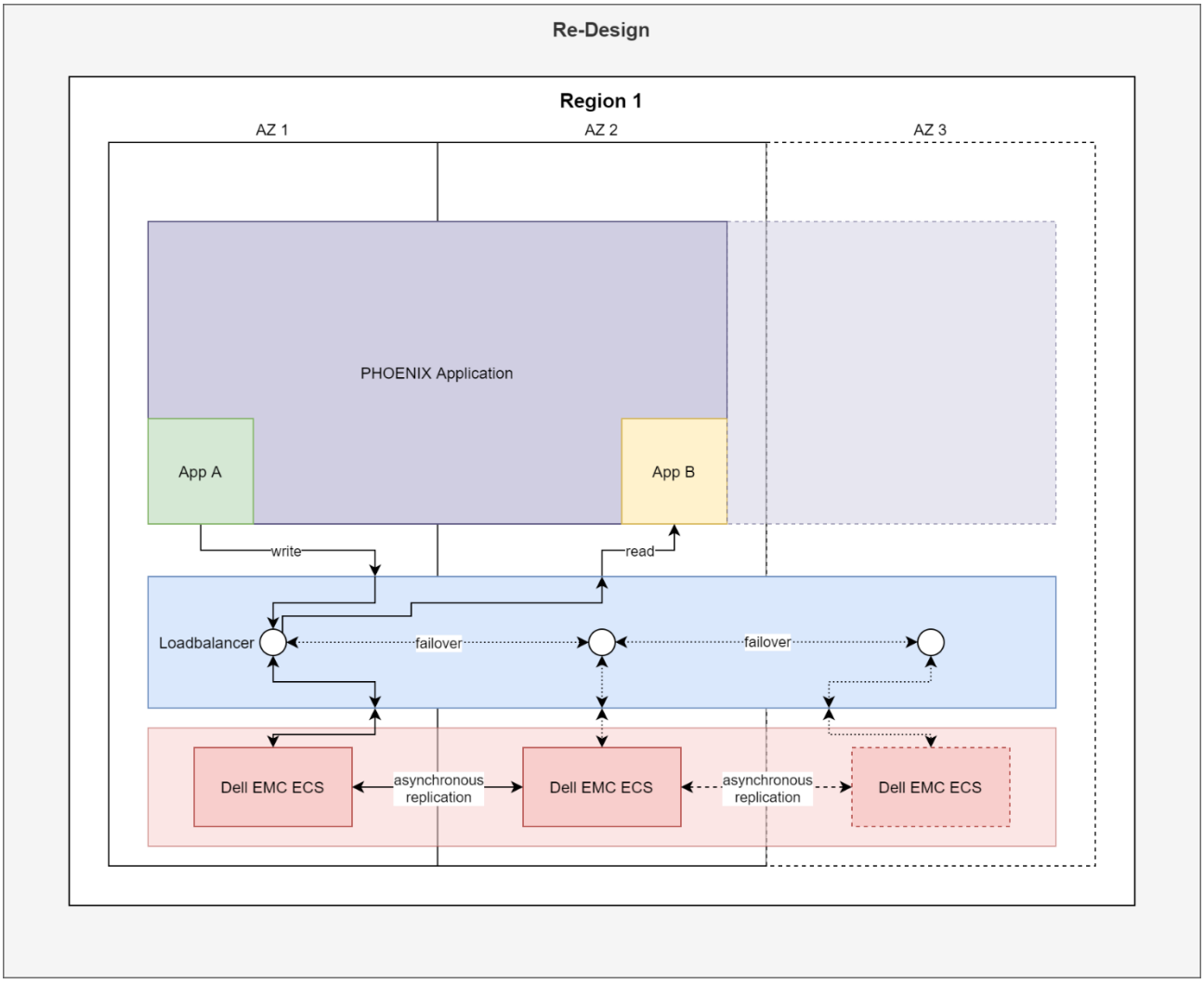

Abbildung 7: Vorschlag für Dell EMC ECS im neuen Design (geplante Elemente sind mit gestrichelten Linien versehen)

Im Vorschlag von Rohde & Schwarz kann das neue Design wie geplant vollzogen werden. Es werden auch weiterhin mehrere *Dell EMC ECS Instanzen* (in Abbildung 7 rot markiert) zusammengeschaltet. Jedoch wird diesen ein *Loadbalancer* (in Abbildung 7 blau markiert) vorgeschaltet. Dieser leitet alle Schreib- und Lesezugriffe auf eine spezifische *Dell EMC ECS Primärinstanz* (z.B. die erste aus Verfügbarkeitszone 1) weiter. Intern repliziert diese *Dell EMC ECS Primärinstanz* ihre Daten, damit im Fehlerfall Sicherungskopien vorliegen. Durch dieses Design wird die inhärente Konsistenz vom aktuellen Design umgesetzt, da beide Applikationen A und B (in Abbildung 7 grün und gelb markiert) auf die gleiche *Dell EMC ECS Primärinstanz* zugreifen. Dies ist komplett unabhängig davon, in welcher Verfügbarkeitszone die Applikationen laufen.

ROHDE&SCHWARZ

Im Fehlerfall (*Dell EMC ECS Primärinstanz* fällt aus) würde der *Loadbalancer* alle Anfragen auf eine andere Instanz umleiten (z.B. die zweite Instanz in Verfügbarkeitszone 2). Nun können Lese- und Schreibzugriffe von Seiten der Applikation ohne Unterbrechung weiter durchgeführt werden. Dadurch bedarf es keiner applikatorischen Änderung und Zugriffe erfolgen aus Applikationssicht auf denselben Objektspeicher. Wie viele *Dell EMC ECS Instanzen* im Hintergrund voll funktionsfähig oder aber ausgefallen sind, weiß die Applikation nicht.

Der einzige Datenverlust, der auftreten könnte, wäre ein Ausfall der *Dell EMC ECS Primärinstanz* während der Replikation. Alle bereits replizierten Daten wären gesichert und weiterhin verfügbar, jedoch die gerade replizierenden Daten nicht. Hier müsste eine Änderung auf Seiten der PHOENIX Applikation vorgenommen werden um diesen Datensatz erneut einzulesen und zu verarbeiten.

7.9 Dell EMC ECS S3 SLA

Verfügbarkeiten werden seitens PSP genauso beim neuen Design zugesichert wie bei dem aktuellen Design. Hier werden keine Änderungen erwartet.

Jedoch muss je nach gewähltem Dell EMC ECS-Design ein temporärer Ausfall betrachtet werden. Im aktuellen Design würde der Ausfall der Dell EMC ECS-Instanz einer Region zum Komplettausfall der Region führen. Die anderen Regionen würden aber weiterarbeiten können, da diese komplett autark sind.

Das neue Design würde einem Multisite-ECS-Deployment entsprechen. Hier kann ein sogenanntes Access During Outage (ADO) Feature genutzt werden. Dabei kann auf Daten zugegriffen werden, wenn ein Standort (in diesem Fall eine Dell EMC ECS Instanz) ausfallen würde (Temporary Site Outage). Das ADO Feature wird auf Bucketebene konfiguriert.

Somit könnte auf Buckets während eines temporären Ausfalls, von einer anderen Instanz aus lesend zugegriffen werden. Dabei ist zu beachten, dass bei Netzwerkausfall diese Lesezugriffe keine Änderungen auf Seiten des Originalbuckets beinhalten, die während des Ausfalls auftreten.

7.10 Auswirkungen des Ausfalls eines Geo-Standortes auf die Verfügbarkeit der Fachanwendung

Das aktuelle Design der Fachanwendung PHOENIX geht von zwei unabhängigen Regionen am Standort Wiesbaden aus. Für beide Regionen ist ein unabhängiges Update sowohl technisch notwendig als auch sinnvoll um eine erweiterte Servicegarantie der Fachanwendung im Fall des Updates einer einzelnen Region zu gewährleisten.

Nach neuem Design existiert zum Zeitpunkt der geplanten Wirkbetriebsaufnahme von PHOENIX zunächst nur eine Region. Aufwände zur Durchführung von Updates der Fachanwendung am Standort Wiesbaden reduzieren sich dann im Vergleich zum aktuellen Design, da nur noch eine einzelne Region betroffen ist. Bis zur Verfügbarkeit der zweiten Region erhöht sich zunächst das Risiko, daß durch Fehler in bzw. Konfigurationsfehler der Fachanwendung, dass die Fachanwendung nicht im erwarteten Umfang verfügbar ist, weil im Fehlerfall nicht auf eine zweite Region ausgewichen werden kann.

Diesem Risiko kann nach Einschätzung von Rohde & Schwarz durch die vorgesehenen mehrstufigen Test- bzw. Abnahmeprozesse entgegengewirkt werden, in dem die Fachanwendung folgende Qualitätsgates durchläuft:

1. Release und Freigabe innerhalb der Entwicklung einer Komponente,
2. Release und Freigabe der Fachanwendung auf dem Entwicklungssystem von Rohde & Schwarz im Rahmen des FAT,

ROHDE&SCHWARZ

3. Release und Freigabe der Fachanwendung auf dem Auftragnehmer-Referenzsystem im Rahmen des Remote-FAT,
4. Release und Freigabe der Fachanwendung auf dem Pre-Produktionssystem des Auftraggebers im Rahmen des SAT

Erst nachdem diese Stufen erfolgreich durchlaufen wurden, wird die Fachanwendung auf dem Produktivstem innerhalb eines definierten Wartungsfensters mittels Blue-Green-Deployment so ausgerollt, dass es zu keiner Serviceunterbrechung der Fachanwendung innerhalb einer Region kommt.

Sollten Fehler in der Anwendung trotz dieser mehrstufigen Abnahmeprozedur nicht gefunden werden, ist es aus Sicht von Rohde & Schwarz sehr unwahrscheinlich, dass diese dann zeitnah auf dem Produktionssystem auftreten und zu einem Ausfall der Fachanwendung führen.

Den zweiten zu betrachtendem Aspekt zu stellen Konfigurationsfehler dar. Da die Regionen, aufgrund ihrer Unabhängigkeit zueinander eigenständige Konfigurationssets besitzen ist aus Sicht der Autoren eine Aufdeckung eines Fehlers im Bereich des Konfigurationsmanagements der Fachanwendung ebenfalls sehr unwahrscheinlich.

Basierend auf dieser Diskussion hat der Wegfall der zweiten Region am Standort Wiesbaden bezüglich der Verfügbarkeit der Fachanwendung in Update bzw. Upgrade Szenarien, aus Sicht von Rohde & Schwarz keine praktische Relevanz und ist somit risikoneutral.

7.11 Auswirkungen auf die persistierenden Datenhaltungskomponenten

Die grundlegende Philosophie bei allen von der Fachanwendung PHOENIX benötigten verteilten Datenhaltungskomponenten (Apache Cassandra, Apache Kafka und Elasticsearch) liegt in der Partitionierung und Replikation. Grundsätzlich werden alle Daten zerteilt, kopiert und anschließend auf Clusterknoten verteilt. Dadurch kann ein Ausfall von einer gewissen Anzahl von Clusterknoten ohne Daten- oder Funktionalitätsverlust gewährleistet werden. Im Folgenden werden die einzelnen Datenhaltungskomponenten und ihre Umsetzung dieser Philosophie näher beschrieben.

7.11.1 Apache Cassandra

Apache Cassandra ist ein verteiltes NoSQL-Datenbankverwaltungssystem[9] für große unstrukturierte Datenbanken. Apache Cassandra ist auf hohe Skalierbarkeit und Ausfallsicherheit ausgelegt, was es ideal für den Einsatz in der Fachanwendung PHOENIX macht. Vergleichbar mit herkömmlichen relationalen Datenbanksystemen speichert es Daten in Tabellen bestehend aus Spalten und Zeilen.

Grundsätzlich verteilt Apache Cassandra Daten auf Knoten eines Clusters. Dazu werden die Daten in sogenannte Partitionen zerlegt, welche eine Repräsentation von Zeilen einer Tabelle sind. Diese Partitionen werden gemäß eines Replikationsfaktors kopiert. Hierbei bedeutet der Replikationsfaktor 3, dass insgesamt 3 Kopien einer jeden Partition existieren. Diese Partitionen werden nun so auf den Clusterknoten verteilt, dass maximale Datensicherheit gewährleistet werden kann. Das bedeutet, dass mehrere Knoten oder gar eine ganze Verfügbarkeitszone ausfallen kann, ohne dass Datenverlust entsteht.

Wie in der Abbildung 8 gezeigt, sind Daten (durch *Data A* und *Data B* gekennzeichnet) auf den verschiedenen Clusterknoten verteilt. Wenn man von einem Komplettausfall der Verfügbarkeitszone 2 (in Abbildung 8 orange markiert) ausgeht, sind weiterhin alle Daten (*Data A* und *Data B*) verfügbar und vom Client erreichbar.

[9] Database Management System, DBMS

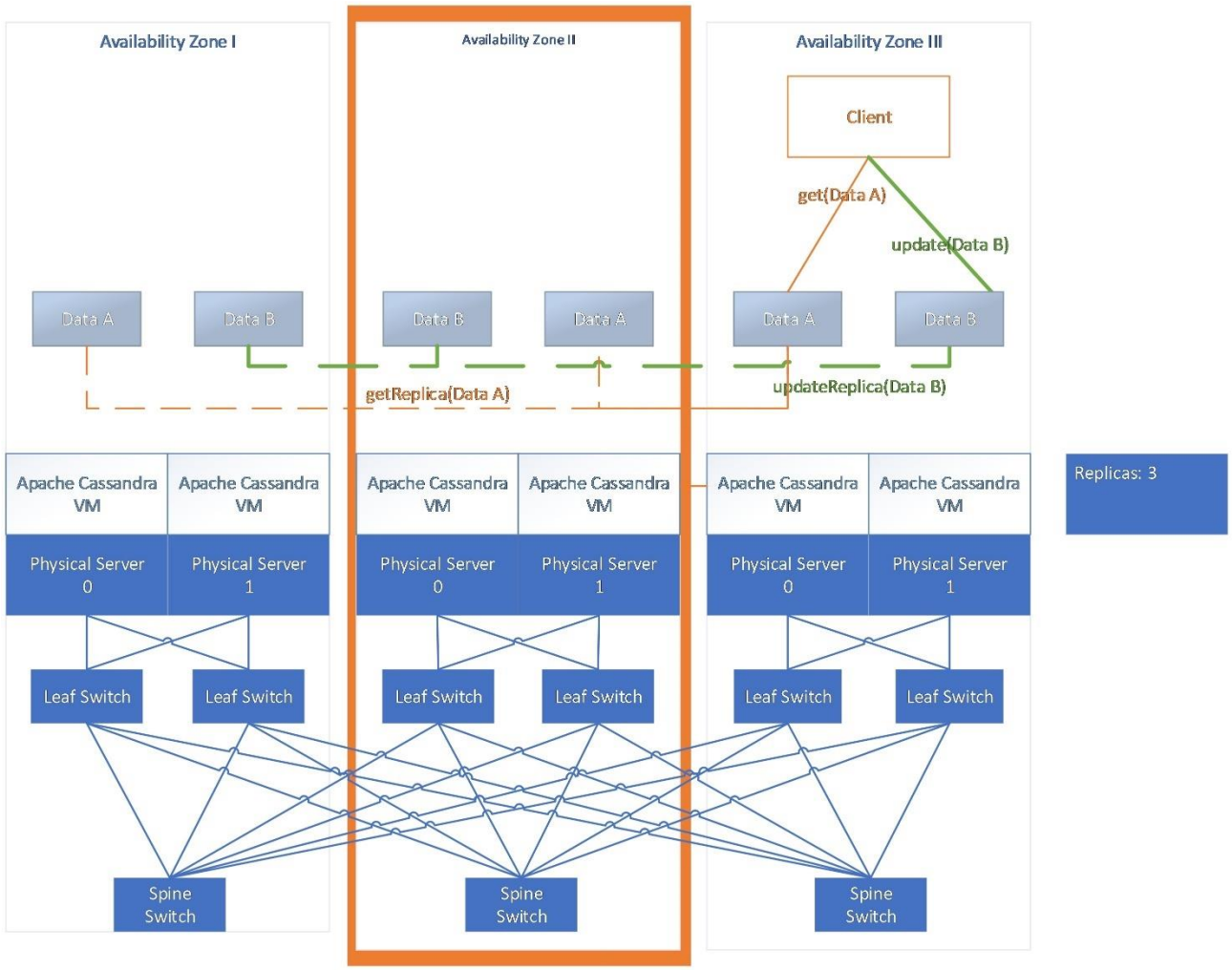

Abbildung 8: Apache Cassandra Schema im neuen Design

7.11.2 Kafka

Apache Kafka ist eine verteilte Plattform um Ereignisdaten und Datenströme zu verarbeiten. Dabei werden sogenannte Topics als logische Konstrukte für die Datenverarbeitung erzeugt. Diese werden in sogenannte Partitions aufgeteilt (nicht zu verwechseln mit Partitionen in Apache Cassandra). Partitions werden zum einen genutzt, um zu skalieren und zum Zweiten, um Datenkopien auf Clusterknoten zu verteilen.

In Apache Kafka wird ein Clusterknoten als Broker bezeichnet. Bei der Erstellung eines Topics mit Partitionierungsfaktor 2, werden zwei Partitions erzeugt (in Abbildung 9 orange und grün markiert). Diese Partitions werden nun gemäß Replikationsfaktor kopiert. Ein Replikationsfaktor von 3 bedeutet, dass drei Kopien einer jeden Partition existieren. Diese Partitions werden nun auf die verschiedenen Broker im Cluster verteilt.

Eine Partition wird dabei als Leader bestimmt, was bedeutet, dass nur von dieser Partition gelesen und geschrieben werden kann. Nach dem Schreiben von Daten auf die Leaderpartition werden diese Daten auf die anderen Replicas repliziert. Dazu benötigt man sogenannte In-Sync Replicas (ISR). Diese Replicas dienen als Konsistenzgarantie. Werden Daten in eine Leaderpartition geschrieben, müssen alle ISRs das Replizieren bestätigen. Anschließend wird auf alle andere Replicas repliziert.

ROHDE&SCHWARZ

Fällt nun ein Broker oder eine ganze Verfügbarkeitszone aus, befinden sich Datenkopien und Replicapartitions auf anderen Brokern des Clusters. Sollte ein Broker mit einer Leaderpartition ausfallen, wird einer der ISRs automatisch neue Leaderpartition und kann somit Lese- und Schreibzugriffe annehmen.

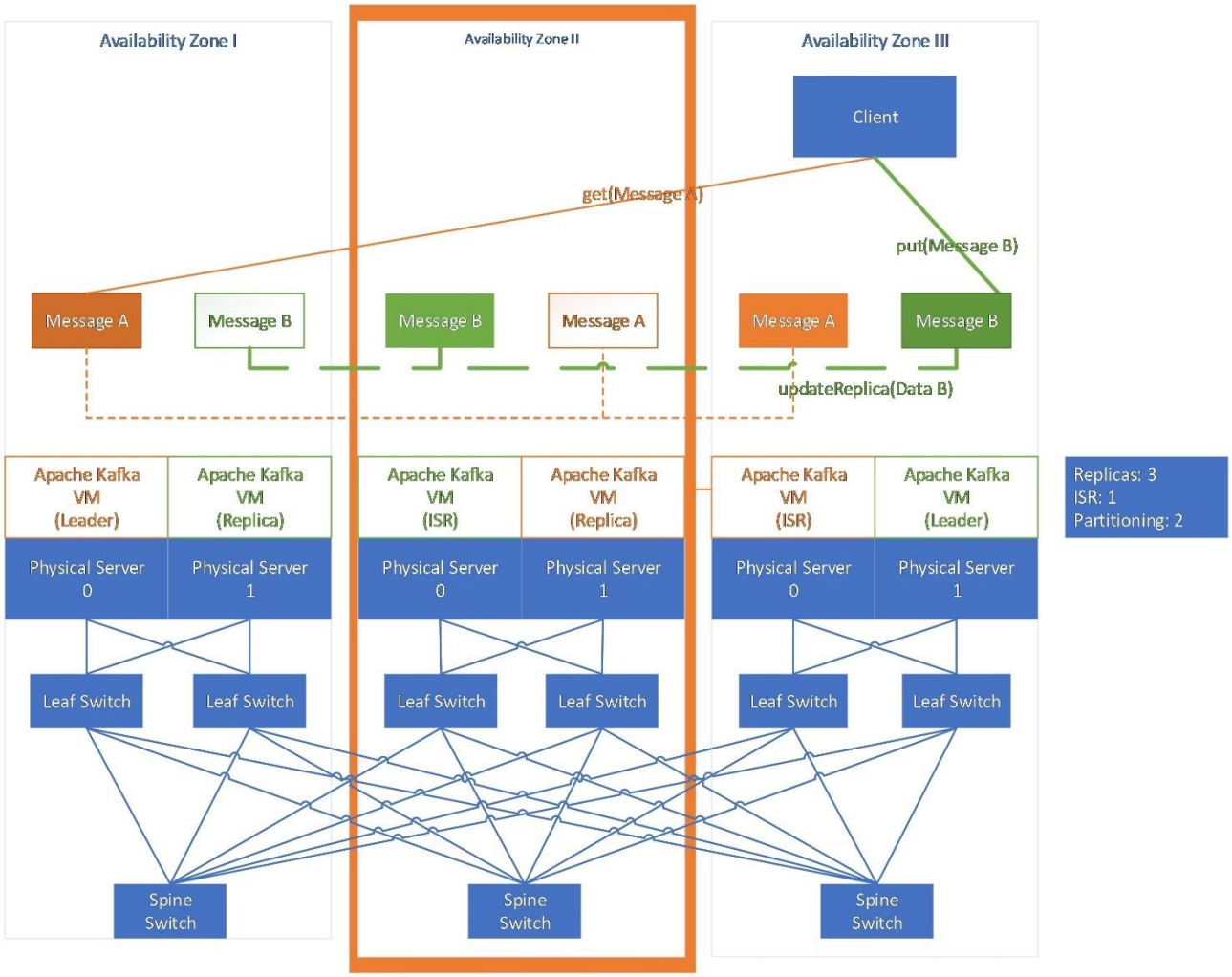

Abbildung 9: Apache Kafka Schema im neuen Design

7.11.3 Elasticsearch

Elasticsearch ist eine Suchmaschine auf Basis von Apache Lucene. Elasticsearch benutzt ähnlich wie Apache Kafka und Apache Cassandra das Prinzip von Partitionierung und Replikation.

Dabei wird ein logisches Konstrukt namens Index erstellt. Indizes sind vergleichbar zu Datenbanken in herkömmlichen DBMS. Sie beinhalten sogenannte Dokumente, welche die eigentlichen Daten beinhalten. Um Indizes zu verteilen werden sie und die beinhalteten Dokumente horizontal in sogenannte Shards[10] aufgeteilt. Diese Shards können auf Clusterknoten verteilt werden und dienen der Skalierung und Datensicherung.

Dabei entstehen bei der Erstellung eines Index eine vordefinierte Anzahl an Shards. Diese initialen Shards werden als Primary Shards bezeichnet. Diese Primary Shards speichern die ursprünglichen Dokumente und somit die Daten. Über den Replikationsfaktor kann nun bestimmt werden, wie viele Kopien dieser Primary Shards existieren. Ein Replikationsfaktor von 1 bedeutet, dass zwei Kopien der Daten existieren:

[10] entspricht einem Apache Lucene-Index

die ursprüngliche Kopie in den Primary Shards und eine zweite Kopie in einem Replica Shard. Erfolgen nun Schreibzugriffe (z.B. Indexierung eines neuen Dokuments), werden diese direkt zu dem Primary Shard geleitet. Das Primary Shard ist sowohl für die eigentliche Indexierung als auch die Replikation verantwortlich. Haben alle Replikate die Änderung übernommen, wird dem Client eine erfolgreiche Transaktion gemeldet. Dadurch ist Konsistenz und Datensicherheit gewährleistet.

Fällt ein Clusterknoten oder eine ganze Verfügbarkeitszone aus (in Abbildung 10 orange markiert) kann der Client weiterhin auf die Daten zugreifen. Dies ist möglich, da die Daten auf andere Shards aufgeteilt und repliziert wurden. Fällt ein Primary Shard aus, wird ein Replikat zum neuen Primary Shard befördert, welches nun wieder Schreibzugriffe annehmen kann. Lesezugriffe können stets von allen Replikas angenommen werden.

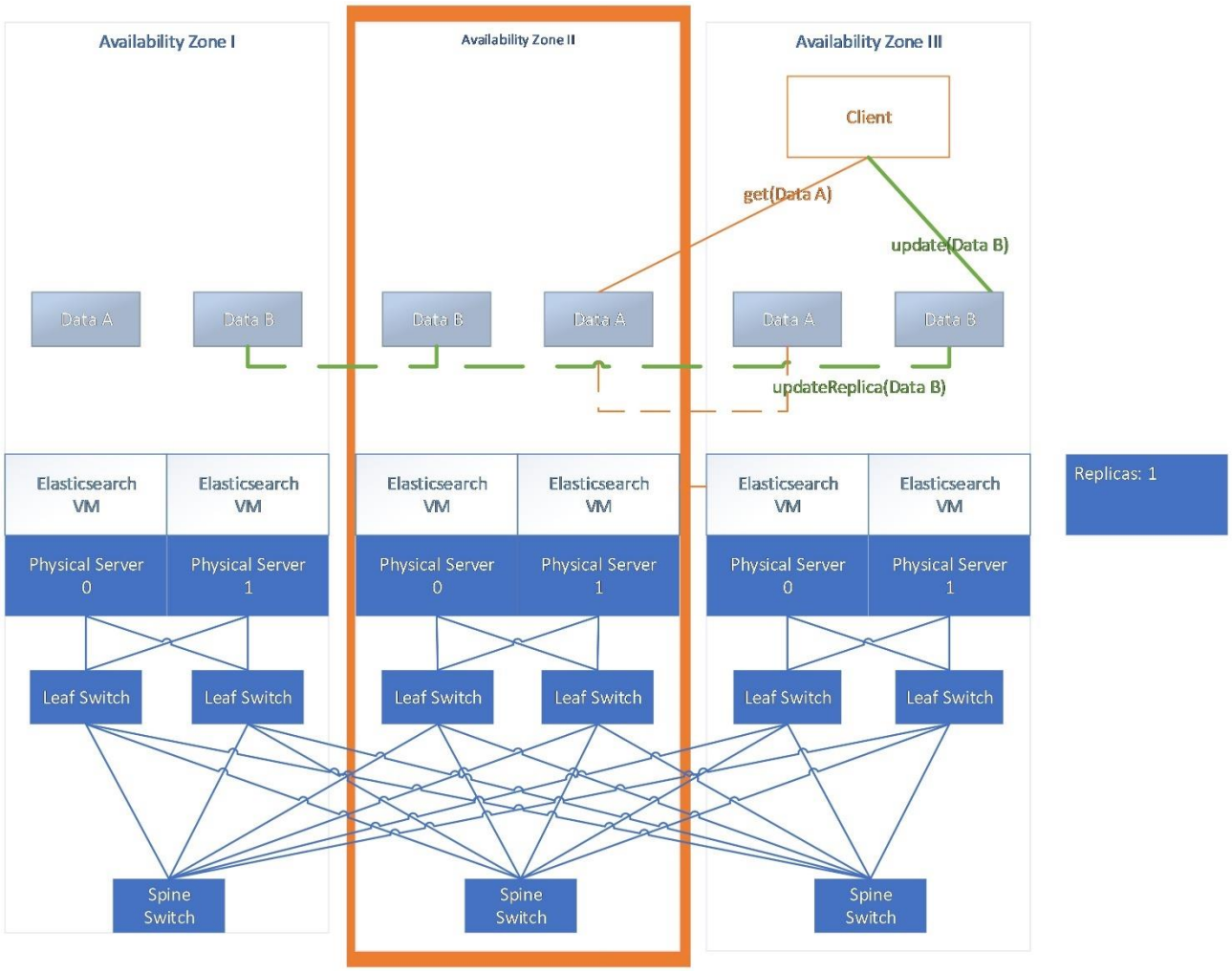

Abbildung 10: Elasticsearch Schema im neuen Design

7.11.4 PostgreSQL

PostgreSQL kann im neuen Design in Abhängigkeit des Betriebsmodells sowohl als Legacy-Service als auch als eine Art Cloud Native Services betrachtet werden. In Bezug auf die Bereitstellung des PostgreSQL-Services innerhalb einer Region sind nach Einschätzung von Rohde & Schwarz und basierend auf den gegebenen Servicegarantien der PSP keine negativen Auswirkungen auf diesen Service zu erwarten.

ROHDE&SCHWARZ

7.11.5 Zusammenfassung

Wie oben beschrieben sind die verteilten Datenhaltungskomponenten so ausgewählt, dass diese unabhängig vom Design in Verfügbarkeitszonen arbeiten können. Die Datenhaltungskomponenten sind außerdem so für die Fachanwendung PHOENIX gestaltet, dass sie mit einem Komplettausfall einer gesamten Verfügbarkeitszone umgehen können. Dies ist in Abbildung 8, Abbildung 9 und Abbildung 10 durch die orange markierte Verfügbarkeitszone dargestellt. Ein Zugriff auf alle Daten ist trotz Ausfall dieser Verfügbarkeitszone weiterhin möglich.

Die Annahme bei den Datenhaltungskomponenten ist, dass Netzwerklatenzen und Bandbreiten gemäß PSP im neuen Design gehalten werden können. Zudem wird angenommen, dass wie von der PSP bestätigt, eine Vollvermaschung aller Verfügbarkeitszonen besteht. Das heißt, dass bei Ausfall einer Verfügbarkeitszone die anderen noch miteinander kommunizieren können. Ist dies der Fall, hat das neue Design keinen Einfluss auf die Datenhaltungskomponenten. Somit besteht kein Risiko in diesem Bereich.

7.12 Diskussion der Reduktion der Komplexität

Durch das neue Design wird das jetzige Konzept von zwei Regionen am Standort Wiesbaden mit jeweils drei Verfügbarkeitszonen auf eine Region am Standort Wiesbaden mit drei Verfügbarkeitszonen[11] reduziert. Diese Reduktion hat Auswirkungen auf die Komplexität des Betriebs der Fachanwendung welche im Folgenden analysiert werden.

7.12.1 Reduktion der Investitionskosten

Durch die Reduktion der Verfügbarkeitszonen am Standort Wiesbaden ist es möglich die verfügbare Hardware effektiver zu nutzen, da keine zwei voll ausgebauten Regionen mit je drei Verfügbarkeitszonen bestückt werden müssen. Dies resultiert in einer Verdoppelung der Hardware-Ressourcen in der einzelnen neuen Region in Wiesbaden. Hierdurch können Kosten für neue Hardware, Software und Lizenzen zur Vergrößerung eingespart werden.

7.12.2 Reduktion der Komplexität in der Fachanwendung

Durch das neue Design wird die Komplexität für die Fachanwendung deutlich reduziert. Im Bereich Replikation kann durch das neue Design auf Mechanismen der Datenhaltungskomponenten zurückgegriffen werden. Dies verringert das Risiko von Ausfällen durch Fehlkonfigurationen und logische Fehler.

Aus unserer Sicht ist die Reduktion der Komplexität ein Vorteil für das Projekt.

7.13 Verfügbarkeitsbetrachtungen

Laut PSP werden die theoretischen Verfügbarkeiten des aktuellen Designs auch beim neuen Design gegeben sein.

Somit besteht kein Risiko in diesem Bereich.

[11] Bezieht sich auf den geplanten Endausbau mit der 3. Verfügbarkeitszone in einem externen Rechenzentrum in synchroner Reichweite.