Bundesamt
für Sicherheit in der
Informationstechnik

Redundanz
Modularität
Skalierbarkeit

Bundesamt für Sicherheit in der Informationstechnik
Postfach 20 03 63
53133 Bonn

E-Mail: gpreferat-b31@bsi.bund.de
Internet: https://www.bsi.bund.de

Inhaltsverzeichnis

1 Einführung

Das bewährteste und bekannteste Mittel zur Sicherstellung der Verfügbarkeit technischer Einrichtungen ist die Redundanz. Im einfachsten Fall ist damit gemeint, dass einem erforderlichen System ein weiteres zur Seite gestellt wird, das bei Ausfall des ersten Systems dessen Funktion übernimmt.

In einem solchen Fall werden somit zwei Systeme vorgehalten, obwohl für den normalen Betrieb ein System reichen würde. Allein dieses einfache Modell zeigt neben dem Nutzen der verbesserten Verfügbarkeit zugleich den Nachteil von Redundanz auf: Redundanz ist unausweichlich mit der Bereitstellung von Überkapazität verbunden.

Mit der Methode der Modularität kann Überkapazität auf ein vertretbares Maß reduziert werden.

Die Skalierbarkeit ist neben Redundanz und Modularität das dritte Design-Merkmal hochverfügbarer Systeme. Sie ermöglicht eine weitestgehend unterbrechungsfreie Kapazitätserhöhung, z. B. bei steigendem Bedarf.

Zweck dieses Dokuments ist es, Beteiligten an Gesprächen über Redundanz, Modularität und Skalierbarkeit eine gemeinsame Verständnisgrundlage zu geben. Die enthaltenen Beispiele sind zwar weitgehend aus dem Bereich der baulichen Infrastruktur entnommen, sie lassen sich aber ohne weiteres auf viele andere Bereiche der IT sinnvoll übertragen.

2 Redundanz

Redundanz ist die Bereitstellung zusätzlicher, über den eigentlichen Bedarf hinaus gehender Systeme. Da es diverse Möglichkeiten gibt, zusätzliche Systeme bereit zu halten, werden im Folgenden die wesentlichen Redundanzformen dargestellt.

2.1 Betriebsredundanz

Die einfachste Redundanz ist die „Betriebsredundanz“. Sie stellt einem System, dessen Leistungsvermögen 100 % beträgt, ein zweites System zur Seite, dessen Leistungsvermögen ebenfalls 100 % beträgt. Fällt das erste System aus, kann das zweite die erforderliche Leistung in vollem Umfang erbringen. Die Bezeichnung „Betriebsredundanz“ ist gewählt, weil durch sie im Falle eines einzelnen Fehlers der Betrieb aufrechterhalten wird.

Diese Redundanz-Form wird auch (N+1)-Redundanz genannt. N ist immer die Anzahl der Systeme, die im normalen Betrieb die erforderliche Leistung bereitstellen. In Bild 2-1 ist der Wert von N also 1.

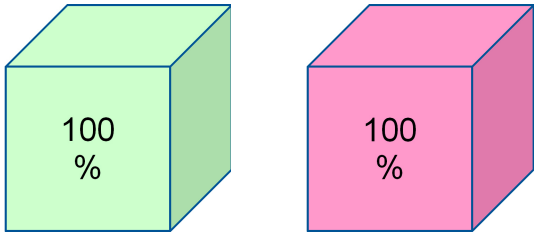

Bild 2-1: Betriebsredundanz (N+1)

Die Betriebsredundanz hat ihre Grenze, wenn eines der beiden Systeme, z. B. zu Inspektions- oder Wartungszwecken außer Betrieb genommen wird. Soll in einen solchen Fall ebenfalls Redundanz gewärleistet sein, ist der Aufbau einer sog. „Wartungsredundanz“ erforderlich.

2.2 Wartungsredundanz

Bei der Wartungsredundanz werden dem primär erforderlichen System zwei gleichwertige leistungsfähige Systeme zugeordnet. Diese Redundanzform wird mit (N+2) bezeichnet, wobei N (im nachfolgenden Bild 2-2) weiterhin den Wert 1 hat.

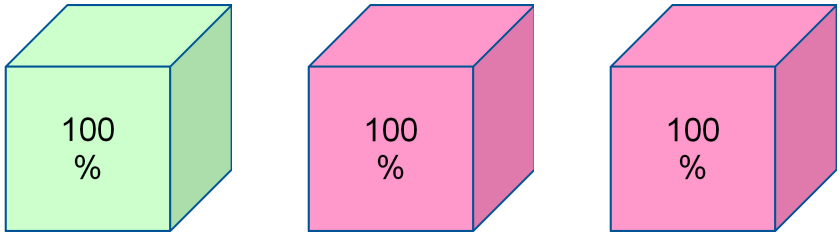

Bild 2-2: Wartungsredundanz (N+2)

Mit dieser Wartungsredundanz ist bei wartungsbedingtem Ausfall eines der drei Systeme die Betriebsredundanz durch die beiden verbleibenden Systeme gegeben. Sie ist also für sehr hohen Anspruch an die Verfügbarkeit grundsätzlich unverzichtbar.

Da diese Redundanzform eine Überkapazität von 200 % mit sich bringt, stößt sie rasch an räumliche und finanzielle Grenzen.

Um das Problem der Überkapazität zu reduzieren, kommt der Mechanismus der Modularität (siehe Kapitel 3 ab Seite 11) ins Spiel. Zunächst müssen aber noch einige Redundanz-Begriffe erläutert werden.

2.3 Tatsächliche Redundanz

Eine wirksame und nutzbare Redundanz ist nur gegeben, wenn die Systeme so angeordnet und miteinander verschaltet sind, dass jedes einzelne ohne Beeinträchtigung anderer Systeme gewartet, repariert und ausgetauscht werden kann. Am Beispiel einer Kühlmittelpumpe wird diese „tatsächliche Redundanz“ erklärt.

Bei einer Zwillingspumpe (Bild 2-3) sind zwar zwei Pumpen vorhanden, beide sitzen aber in einem gemeinsamen Grundgehäuse. Muss eine der beiden Pumpen ausgetauscht werden, müssen die Schieber (im Foto rot) über und unter dem Pumpengehäuse geschlossen werden. Obwohl zwei Pumpen vorhanden sind, steht bei einer Reparatur keine zur Verfügung.

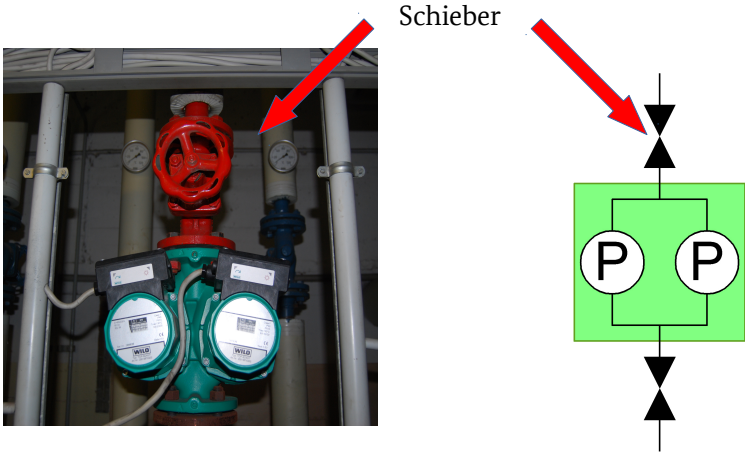

Bild 2-3: Zwillingspumpe

Anders ist das bei einer Doppelpumpe (Bild 2-4) aus. Dabei sind die beiden Pumpen als jeweils selbstständige Einheit aufgebaut.

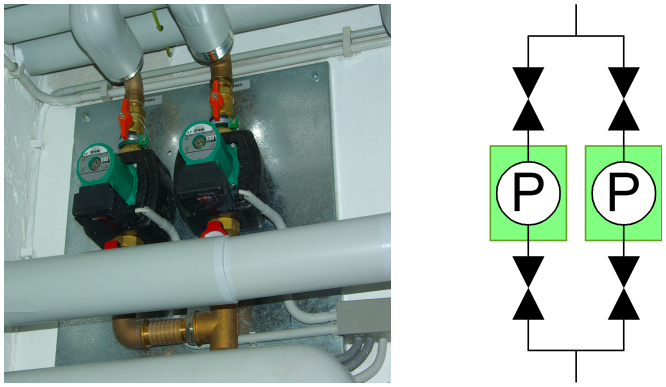

Bild 2-4: Doppelpumpe

Durch Schließen der Schieber des jeweiligen Pumpenstrangs kann jede Pumpe ohne Beeinträchtigung der Funktionsfähigkeit der anderen repariert werden. Voraussetzung ist, dass die Pumpen so dimensioniert sind, dass die Förderleistung jeder einzelnen ausreichend ist, also die benötigte Leistung von 100 % liefern kann.

Sind die Pumpen hingegen so dimensioniert, dass sie nur gemeinsam die erforderliche Förderleistung erbringen, ist das in keiner Weise eine Redundanz, sondern lediglich eine Lastverteilung. Auch bei einer solchen Pumpenanordnung wird der Ausfall einer Pumpe zu Problemen führen, weil die benötigte Leistung nicht mehr zur Verfügung steht.

2.4 Temporäre Redundanz

Mitunter scheitert der permanente Aufbau einer Wartungsredundanz an den Kosten oder dem nicht verfügbaren Platz. Hier kann das Mittel der temporären Redundanz Abhilfe schaffen. Dabei wird als Dauerinstallation eine (N+1)-Betriebsredundanz aufgebaut. Steht eine geplante Abschaltung für eine Inspektion oder Wartung an, kann im Vorfeld eine dritte Einheit hinzu geschaltet werden, sodass auch während der Wartung die Betriebsredundanz sichergestellt ist.

Wichtig dabei ist, alles für die Inbetriebnahme der temporäre Wartungsredundanz so vorzubereiten, dass dadurch zu keiner Beeinträchtigung der vorhandenen Redundanz führt.

Bei Netzersatzanlagen (NEA) wird z. B. neben zwei ortsfesten Anlagen ($oNEA_1$ und $oNEA_2$), welche die normale (N+1)-Redundanz liefern, ein Anschlusspunkt für eine dritte, mobile Einheit mNEA vorbereitet. Die mNEA kann bei Bedarf ins System eingebunden und in Betrieb genommen werden, ohne dass dadurch die Funktionen der ortsfesten Anlagen beeinträchtigt werden.

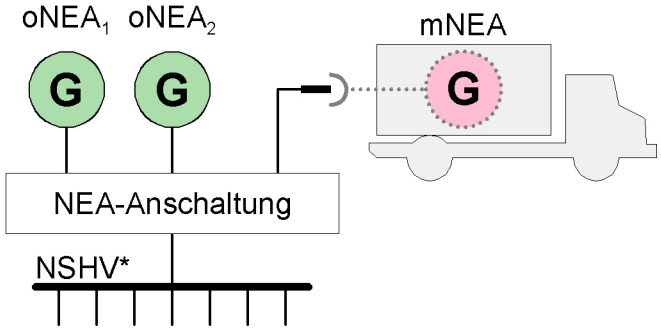

Bild 2-5: Vorbereitung für eine temporäre (N+2)-Redundanz
(* = Niederspannungshauptverteilung)

2.5 Redundanz durch Diversität

Diversität bezeichnet eine Redundanz, die auf funktional gleichwertigen, aber unterschiedlich implementierten Komponenten basiert. Diversitär redundante Komponenten tragen in gleicher Weise zur Erbringung der Nutzfunktionen in einer hochverfügbaren Architektur bei, basieren aber auf unterschiedlichen Varianten der Realisierung.

Für den Bereich der baulichen und technischen Maßnahmen zum Schutz der IT-Sicherheit bedeutet das, Systeme zu nutzen, die sich in ihrer Funktion gleichen, aber von unterschiedlichen Herstellern stammen.

Die Anwendung der Redundanz durch Diversität muss sorgsam abgewogen werden, da sie dem einfachen Aufbau von Modularität und Skalierbarkeit entgegenstehen kann und ggf. den Betrieb und die Wartung der Systeme komplizierter macht.

Bei sehr hohen Anforderungen an die Verfügbarkeit sollte für komplexe elektromechanische sowie elektronisch gesteuerte Komponenten (USV[1], Transferschalter, Klimatechnik usw.) auf Diversität hingegen nicht verzichtet werden, z. B. durch die Verwendung von Systemen unterschiedlicher Hersteller in den einzelnen Pfaden mehrpfadiger Strukturen. Bei diesen komplexen Komponenten werden Design- oder Konstruktionsfehler mitunter erst Jahre nach der Markteinführung erkannt. Solche Fehler können dann zum gleichzeitigen Ausfall redundanter Systeme führen.

Bei mechanischen und einfachen elektromechanischen Gerätschaften (Ventilatoren, Pumpen, Stromschienen, einfache Generatoren usw.) kann dann auf Diversität verzichtet werden, wenn

- erprobte und bewährte Komponenten genutzt werden, also auf die Nutzung neuester, möglicherweise noch nicht ausgereifter Technik verzichtet wird.

1 Unterbrechungsfrei Strom-Versorgung

Dieser Verzicht ist akzeptabel, weil davon ausgegangen werden darf, dass derartige Geräte nach der Erprobung beim Hersteller und nach einer mindestens einjährigen Nutzung im Markt eine ausreichende Reife erreicht haben und somit anfängliche Serienfehler eliminiert sind oder

- Ersatzgeräte in ausreichend kurzer Zeit voll betriebsfähig eingebaut und angeschlossen werden können oder
- für Komponenten, soweit das technisch möglich ist, temporäre Redundanz (siehe 2.4) vorbereitet wird.

2.6 Georedundanz

Mit Georedundanz ist gemeint, dass einander Redundanz gebende Systeme räumlich über eine größere Entfernung voneinander aufgebaut sind. Ziel ist es, sicherzustellen, dass auch ein sehr großräumiges Ereignis maximal eines der georedundanten Systeme treffen kann. Die hierzu in diversen Veröffentlichungen genannten Entfernungen reichen je nach Quelle und beabsichtigtem Nutzen der Georedundanz von wenigen 10 Kilometern bis zum Aufbau auf verschiedenen Kontinenten.

2.7 Weitere Redundanzbegriffe

Im Zusammenhang mit der Bereitstellung von Redundanz werden folgende weitere Begriffe verwendet: **passive und aktive Redundanz** sowie **kalte, warme und heiße Redundanz**. Damit wird beschrieben, in welchem Betriebszustand die Redundanz bereitgehalten wird.

- Ist die Redundanz im normalen Betrieb funktionslos und wird erst bei Bedarf aktiviert oder zugeschaltet, wird sie „passive“ oder „kalte“ Redundanz genannt.

  Solche Systeme können in der Institution selber vorrätig gehalten oder über einen externen Dienstleister angemietet werden. Hierzu sind entsprechende SLAs mit dem Dienstleister zu vereinbaren und es müssen die erforderlichen Anschlusspunkte vorbereitet sein. Systeme, die für den Einsatz als temporäre Redundanz (Abschnitt 1.4) bereit gehalten werden, sind eine solche kalte Redundanz.
- Befindet sich ein Redundanz gebendes System in Bereitschaft, benötigt aber für die vollständige Betriebsübernahme eine Aktivierungszeit, ist das eine „warme“ Redundanz.
- Von „aktiver“ Redundanz ist die Rede, wenn Redundanz-Systeme, bei Ausfall eines Systems unterbrechungsfrei den Betrieb übernehmen können.

  Diese Betriebsart darf aber nicht mit einer einfachen Lastverteilung verwechselt werden. Bei der sind ebenfalls zwei oder mehr Systeme an der dauerhaften Leistungserbringung beteiligt, allerdings mit dem Unterschied, dass die erforderliche Leistung nur zur Verfügung steht, wenn alle beteiligten Systeme in Betrieb sind.

Weder sind die vorgenannten Begriffe eindeutig und allgemeinverbindlich definiert, noch ist in irgendeiner Weise festgelegt, ob die Aktivierung redundanter Systeme automatisch oder händisch erfolgt.

2.8 Ringeinspeisung

Die Ringeinspeisung ist keine eigenständige Art von Redundanz, sondern eine Maßnahme Redundanz zu realisieren. Da sie aber im Zusammenhang mit Redundanz der Stromversorgung häufig als Maßnahme zur Erhöhung der Verfügbarkeit genannt wird, was aber so pauschal nicht zutrifft, wird sie hier kurz angesprochen.

Bei einer Ringeinspeisung erfolgt die Versorgung über zwei, räumlich möglichst weit auseinander verlegte Trassen, von denen jede Trasse allein den erforderlichen Versorgungsbedarf decken kann. Das Maß, in dem

durch eine Ringeinspeisung eine Erhöhung der Verfügbarkeit bewirkt wird, ist stark von der konkreten Ausführung der Ringeinspeisung abhängig. Dabei sind die beiden folgenden Aspekte zu betrachten:

- Welcher Abstand besteht zwischen den Stellen, an denen die beiden Stränge des Rings aus dem öffentlichen (ungesicherten) Bereich in den eigenen (gesicherten) Bereich übergehen?

  Eine Ringeinspeisung verliert viel von ihrer positiven Wirkung für die Verfügbarkeit, wenn die letzten Meter im öffentlichen Gelände so nah beieinander verlegt sind, dass sie bei Tiefbauarbeiten gleichzeitig oder sehr zeitnah beschädigt werden können. (Stichwort „Baggerrisiko“)

- Entspringen beide das Objekt versorgenden Stränge des Rings aus der gleichen Quelle oder kommen sie aus unterschiedlichen?

  Wenn beide Stränge aus dem gleichen Umspannwerk oder aus der gleichen Schaltanlage heraus geführt sind, stellen diese Komponenten einen SPOF (Single Point Of Failure) dar. Fällt das Umspannwerk oder die Schaltanlage aus, fallen beide Ring-Stränge gleichzeitig aus.

3 Modularität

Modularität beschreibt, ob die erforderliche Leistung durch eine große oder mehrere kleinere parallel arbeitende Einheiten zur Verfügung gestellt wird. Es geht also um den Wert von N.

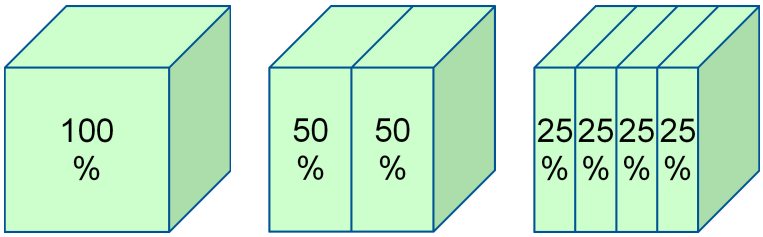

Bild 3-1: Modularität mit N=1, N=2 und N=4

Stellt z. B. eine USV alleine eine erforderliche Leistung von 120 kVA bereit, ist N=1. Wird die Leistung hingegen durch 4 USVen zu je 30 kVA im Parallelbetrieb bereitgestellt, ist N=4.

Die Anwendung der Modularität hat mehrere Vorteile. Der eine ist die Reduzierung von Überkapazität. Der andere Vorteil ist die im Fehlerfall höhere Restkapazität.

3.1 Überkapazität

Bei einer (N+1)-Redundanz mit N=2 sinkt die Überkapazität von 100 % auf 50 %.

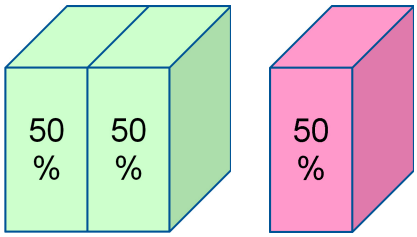

Bild 3-2: (N+1)-Redundanz mit (N=2)

Mit dem gleichen Umfang an Überkapazität, also 50 %, lässt sich bei einem Wert von N=4 sogar eine (N+2)-Redundanz, also eine Betriebs- und Wartungsredundanz realisieren. Mit N=1 war der Wert der Überkapazität hier noch 200 %, Siehe Bild 2-2 auf Seite 8.

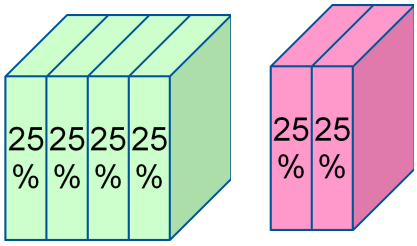

Bild 3-3: (N+2)-Redundanz mit (N=4)

Je höher man den Wert für N treibt, desto geringer wird die Überkapazität, womit deren Kosten sinken. Allerdings hat dieser Weg auch seine Grenzen, denn mit steigendem Wert von N

- steigen die Kosten für die Unterbringung der Einheiten. Es ist dringend zu empfehlen, die Einheiten (im letzten Beispiel sind das schon 6) so voneinander getrennt unterzubringen und zu versorgen, dass durch ein externes Ereignis (z. B. Feuer) keinesfalls alle Einheiten zugleich betroffen sind.
- steigt die Komplexität des Gesamtsystems, was tendenziell zu einer Abnahme der Zuverlässigkeit führen kann.

- steigt durch die zunehmende Zahl von Einzelkomponenten die Wahrscheinlichkeit eines Komponentenausfalls.

In der folgenden Tabelle 1 wird die Gesamt- und die Überkapazitätswerte für steigende Werte von N dargestellt.

N	N+1	Gesamtkapazität	Überkapazität	N+2	Gesamtkapazität	Überkapazität
1	2	200 %	100 %	3	300 %	200 %
2	3	150 %	50 %	4	200 %	100 %
3	4	133 %	33 %	5	167 %	67 %
4	5	125 %	25 %	6	150 %	50 %
5	6	120 %	20 %	7	140 %	40 %
6	7	117 %	17 %	8	133 %	33 %
7	8	114 %	14 %	8	129 %	29 %
8	9	113 %	13 %	10	125 %	25 %

Tabelle 1: Kapazitätswerte für die Redundanz-Modelle (N+1) und (N+2) bei steigenden Werten von N

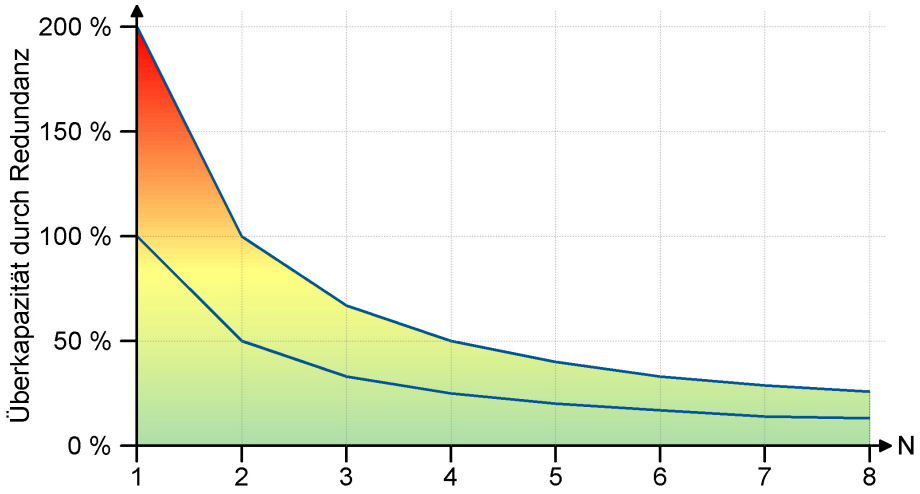

Bild 3-4: Überkapazität durch Redundanz bei steigendem N

In der Grafik in Bild 3-4 lässt sich erkennen, dass schon N=2 einen beträchtlichen Beitrag zur Reduzierung der Überkapazität leistet. Werte für N größer als 4 bringen hingegen nur noch einen so geringen Beitrag, dass die Nachteile durch die steigende Komplexität des Gesamtsystems die Vorteile überdecken.

3.2 Restkapazität

Neben der Reduzierung der Überkapazität hat die Modularität auch noch einen weiteren Vorteil. Sollten mehr Systeme ausfallen, als für das jeweilige Redundanz-Modell angenommen, erhöht die Modularität die verbleibende Restkapazität. Als Beispiel soll hier eine (N+2)-Redundanz mit N=1 und mit N=4 verglichen werden.

Im ersten Fall, also bei (N+2) mit N=1, stehen drei Systeme zur Verfügung. Fallen bis zu zwei Systeme aus, stellt das verbleibende dritte die erforderliche Leistung vollständig bereit. Fällt nun auch noch das dritte System aus, steht KEINE Leistung mehr zur Verfügung.

Im zweiten Fall, also bei (N+2) mit N=4, sieht das schon wesentlich besser aus. Hier stehen insgesamt 6 Systeme zur Verfügung. Bis zu einem Ausfall von 2 Systemen können vom verbleibenden Rest immer noch die erforderlichen 100 % Leistung erbracht werden. Fällt nun ein drittes System aus, sinkt die bereitgestellte Leistung nicht, wie bei N=1, auf 0 % sondern nur auf 75 %.

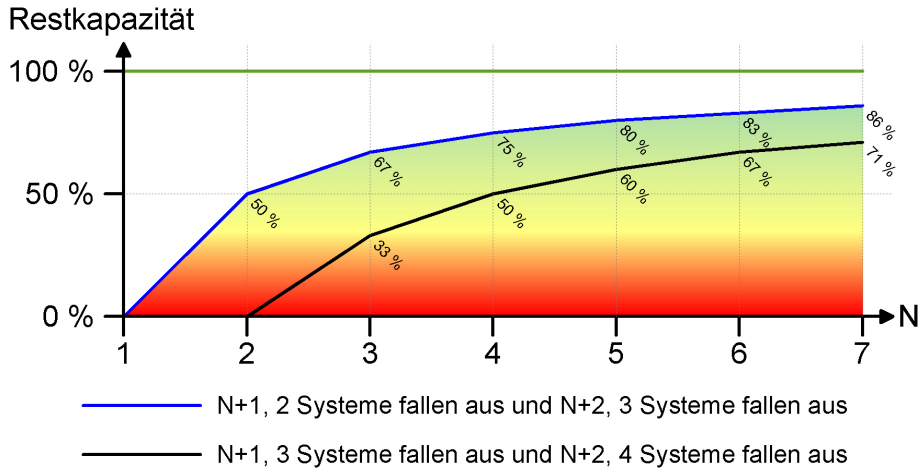

Bild 3-5: Restkapazität für die Redundanzmodelle (N+1) und (N+2) bei Ausfall mehrere Systeme

Bei der Betrachtung der Restkapazität für verschiedene Szenarien (siehe Bild 3-5) zeigt sich, dass auch hier ein Wert für N, der über 4 hinausgeht, nur noch minimalen Mehrnutzen bringt. Gleichzeitig bewirkt ein Wert N=2 schon dann einen deutlichen Nutzen, wenn maximal ein System mehr ausfällt, als eigentlich in dem jeweiligen Redundanzmodell angenommen.

Über die reine Restwert-Betrachtung hinaus kommt noch ein weiterer Aspekt ins Spiel, nämlich die oft nicht vollständige Auslastung der verfügbaren Leistung.

In der Regel liegt die tatsächlich erforderliche Auslastung technischer Systeme unterhalb des 100 %-Maximalwerts. Es ist durchaus zulässig, hier von einem Wert zwischen 60 % und 80 % auszugehen. In vielen Fällen dürften also die im oben genannten Beispiel verbleibenden 75 % Restkapazität vollkommen ausreichen, um alle angeschlossenen Nutzer mit der erforderlichen Leistung zu versorgen.

Selbst wenn das nicht so sein sollte, besteht noch die Möglichkeit, im Rahmen eines Lastmanagements niedriger priorisierter Nutzer in dem Umfang stillzulegen, der erforderlich ist, um mit der Restkapazität die wichtigen Nutzer sicher versorgen zu können.

4 Skalierbarkeit

Keine noch so weitsichtige Planung kann verhindern, dass es nach einiger Zeit erforderlich ist, vorhandene technische Systeme einem geändertem, meist gestiegenem Leistungsbedarf anzupassen. Je einfacher ein System durch simples Hinzufügen zusätzlicher Einheiten erweiterbar ist, desto besser ist es skalierbar.

Meist ist es möglich, bestehende Systeme hinsichtlich der von ihnen erbrachten Leistung nach oben zu erweitern. Wenn diese Leistungserhöhung aber mit der Notwendigkeit verbunden ist, vorhandene Systeme vorübergehend außer Betrieb zu nehmen, sie mit massiven Eingriffen in den Bestand umzubauen oder gar komplett durch neue zu ersetzen, ist keine Skalierbarkeit gegeben.

Skalierbarkeit benötigt zwei ganz wesentliche Eigenschaften:

- Es muss der für die Erweiterung erforderliche Platz zur Verfügung stehen,
- und muss möglich sein, die Erweiterung durch das Hinzuschalten ergänzender Systeme ohne oder durch minimale Betriebsunterbrechungen der Bestandssysteme in Betrieb nehmen zu können.

Am Beispiel einer Elektroverteilung soll das kurz erläutert werden.

Bild 4-1: Elektro-Verteilungen, links und Mitte nicht skalierbar, rechts[2] skalierbar

In Bild 4-1 ist links eine Verteilung gezeigt, die schon im Erstausbau so dicht bestückt wurde, dass es keinerlei Möglichkeit für Erweiterungen gibt.

In der mittleren Verteilung gibt es zwar Platz für Erweiterungen, sie ist aber so aufgebaut, dass diese nur realisiert werden können, wenn die Verteilung, weil aus Arbeitsschutzgründen so vorgeschrieben, komplett spannungsfrei gemacht wird. Damit wären alle daran angeschlossenen Verbraucher für die Dauer der Arbeiten nicht versorgt.

Die Verteilung im rechten Bild ist so aufgebaut, dass Erweiterungen im laufenden Betrieb vorgenommen werden können, ohne dass dazu die bestehenden Versorgungen unterbrochen werden müssen. Diese Verteilung ist also skalierbar.

2 Bildquelle: Werner Henke, Vermögen und Bau Baden-Württemberg, Amt Konstanz
alle anderen Bilder und Zeichnungen: BSI

5 Fazit

Redundanz, Modularität und Skalierbarkeit sind elementare Design-Prinzipien, ohne die hoch- und höchstverfügbare Systeme nicht realisierbar sind. Dies ist natürlich mit zusätzlichen Kosten verbunden. Nicht immer lassen sich alle drei Prinzipien in der hier grundsätzlich dargestellten klar erkennbaren Art und Weise umsetzen.

Sind sie nicht oder nur eingeschränkt umsetzbar gelten die immer gleichen Regeln: Analyse des entstehenden Risikos und Suche nach möglichst wirksamen Alternativen.

Es wird nicht abgestritten, dass durch Redundanz, Modularität und Skalierbarkeit zusätzliche Kosten entstehen, die auf den ersten Blick unwirtschaftlich erscheinen und deren Nutzen, solange nichts passiert, auch nie nachweisbar ist. Es ist aber ebenso unbestritten, dass Redundanz, Modularität und Skalierbarkeit in modernen hochverfügbaren Systemen keine „Zusatzforderungen“ sind, sondern zu den unverzichtbaren Standard-Ausstattungsmerkmalen gehören.