atp magazin

SONDERDRUCK

Transforming Automation
Ausgabe 04 2022

Peer-Review: 21.09.2021

Industrial Edge Cloud für die Smart Factory

Failover und Security für Industrial Edge Computing

Volkan Gezer, Carsten Harms, Deutsches Forschungszentrum für Künstliche Intelligenz; Carsten Brüggemann, Pfalzkom; Michael Pfeifer, Andreas Michael, TÜV Süd; Simon Althoff, Weidmüller; Torsten Runge, Deutsche Telekom/T-Systems; Keran Sivalingam, Technologie-Initiative; SmartFactory KL e.V.; Martin Ruskowski, Deutsches Forschungszentrum für Künstliche Intelligenz

Es gibt mehrere vorgeschlagene Architekturen für das Edge Computing, aber es gibt bislang keine von der Community oder der Industrie akzeptierten Standards. Außerdem gibt es keine gemeinsame Vereinbarung darüber, wie die Edge Computing-Architektur physisch aussieht. In diesem Artikel wird die Industrial Edge Cloud beschrieben, erklärt, wie eine Industrial-Edge-Cloud-Architektur aussieht, welche Anforderungen sie stellt und welche Möglichkeiten sie bietet. Die wichtigsten Funktionen, die ein Edge Node unterstützen sollte, werden ebenfalls definiert. Weiter wird ein Einblick in die Herausforderungen gegeben, die mit der Vernetzung von Maschinen und Anlagen einhergehen und analysiert, warum es nicht ausreicht, lediglich Maschinen oder Anlagen „secure zu machen", sondern dass sich neue Schnittstellen der Cybersecurity mit Betrieb, Instandhaltung, Safety etc. ergeben, die entsprechende Abstimmungsarbeit erfordern.

#Edge Computing #Industrial Edge Cloud #Security at Edge #Offloading #Failover #Recovery

Industrial edge cloud in the smart factories

Failover and security for industrial edge computing

Several architectures have been proposed for edge computing, but so far no standards have been accepted by the community or industry. In addition, there is no common agreement on what the Edge Computing architecture should look like. We describe the Industrial Edge Cloud and explain Industrial Edge Cloud architecture, its requirements and its capabilities. We also define the key features that an Edge Node should support. Furthermore, we give a short insight into the challenges that come along with the networking of machines and plants and discuss why it is not sufficient to simply „secure" machines or plants. New interfaces of cybersecurity with operation, maintenance, safety, etc. require corresponding security work.

#Edge computing #industrial edge cloud #security at edge #offloading #failover #recovery

1. Industrial Edge Cloud

1.1 Grundlagen und Definitionen

Das sogenannte *Computing*, d. h. die Ausführung von Rechenoperationen auf Rechnersystemen, hat seit den 1960er Jahren mehrmals zwischen zentralisierten und dezentralisierten Architekturansätzen gewechselt. In den 1960er Jahren verfügten lokale Endgeräte nicht über ausreichend Rechenleistung, und so wurden die Rechenaufgaben auf zentralen Rechenknoten, den sogenannten *Mainframes*, ausgeführt. Zentralisierte Systeme haben geringere Verwaltungs- und Betriebskosten, und ihre Konfigurationen sind im Vergleich zu den dezentralen Systemen einfacher [1].

Mit der Einführung des Personal Computers (PC) und dessen gestiegener Rechenleistung in den 1980er Jahren konnten Rechenaufgaben direkt auf den lokalen Client-Rechnern exekutiert werden, wodurch die Notwendigkeit zentraler Server zunehmend entfiel [2].

Mit dem zunehmenden Aufkommen der Cloud-Computing-Technolgie in den 2000er Jahren ist wiederum ein Architekturwechsel zu zentralisierten Rechensystemen zu verzeichnen. Cloud Computing kann bezüglich des Bereitstellungsmodells, mit welchem u. a. die betrieblichen Verantwortlichkeiten für die benötigte IT-Infrastruktur definiert werden, in Public, Private and Hybrid Clouds klassifiziert werden.

Bei einer Public Cloud wird die benötigte IT-Infrastruktur von einem Drittanbieter wie beispielsweise Amazon Web Services (AWS) oder Open Telekom Cloud (OTC) in zentralisieren Rechenzentren bereitgestellt und verwaltet, welche häufig geografisch weit vom Endanwender entfernt sind [3]. Die IT-Infrastruktur bei Public Clouds steht dem Endanwender nicht exklusiv zur Verfügung und wird auch durch andere Endanwender parallel genutzt. Durch eine effiziente Auslastung der IT-Infrastrukturen können so Skaleneffekte erzielt werden.

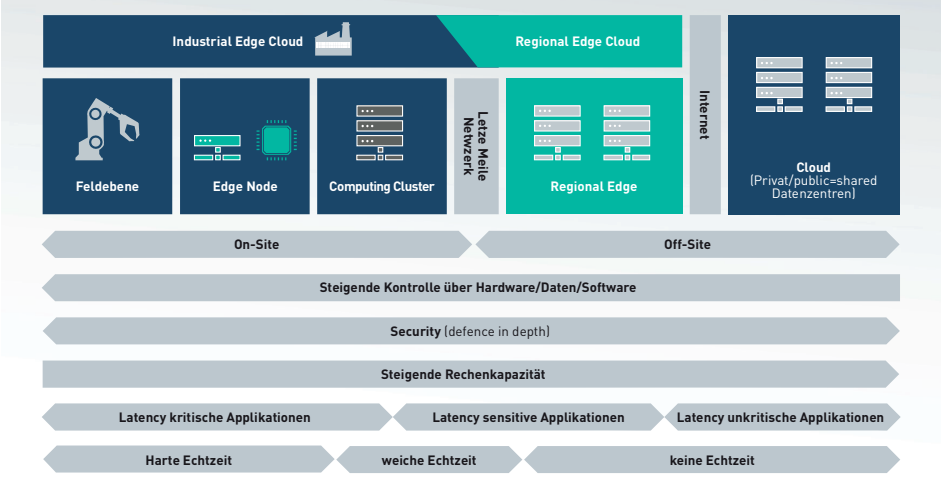

Abbildung 1: Merkmale und Anforderungen von Edge Cloud Instanzen [1].

Im Gegensatz dazu wird bei einer Private Cloud eine dedizierte IT-Infrastruktur für eine eingeschränkte Gruppe von Endbenutzern (z. B. ein bestimmtes Unternehmen) zur Verfügung gestellt, welche nicht von anderen Anwendern genutzt wird. Die Bereitstellung und Verwaltung der IT-Infrastruktur einer Private Cloud kann von dem jeweiligen nutzenden Unternehmen selbst oder alternativ durch einen Drittanbieter zur Auslagerung des IT-Betriebs durchgeführt werden.

Eine Hybrid Cloud stellt eine Mischform zwischen Public und Private Clouds dar, wo beispielsweise aus Datenschutzgründen einzelne Cloud-Services in einer Private Cloud, aber die restlichen Services aus Effizienzgründen in einer Public Cloud ausgeführt werden.

Insbesondere ist für den Zugriff auf eine zentralisierte Cloud eine Internetverbindung notwendig. Aufgrund von Signallaufzeiten sowie der „Best-Effort"-Charakteristik des Internets kann es zu Verzögerungen in der Datenübertragung kommen, was für zeitkritische Applikationen unzureichend sein kann. Um bestimmte Rechenoperation in Echtzeit, also innerhalb einer garantieren Zeit durchzuführen, muss die Ausführung der Rechenoperationen möglichst in der Feld- oder Geräteebene der jeweiligen Anwendung stattfinden. *Edge Computing* (EC) bringt Rechenleistung so nah wie möglich an die Anwendung (z. B. Maschinen oder Werker) heran [4]. Daher wird erwartet, dass EC besonders die Anforderungen von zeitkritischen Applikationen erfüllen kann. Cloud und Edge Computing ergänzen sich miteinander, indem bevorzugt zeitkritische Anwendungen auf Edge-Rechenknoten ausgeführt werden.

Für industrielle Anwendungen im Produktionsumfeld werden Cloud- und Edge Computing aufgrund der Ressourceneffizienz sowie der verbesserten Latenz-Eigenschaften zunehmend interessant. Eine Vereinheitlichung der Taxonomie in Bereich Edge Computing wurde von der Linux Foundation vorgeschlagen [5], welche im Folgenden für Anwendungsgebiete im industriellen Kontext in der Produktion (engl. *shopfloor*) abgeleitet wird. Abbildung 1 veranschaulicht die bevorzugten Grenzen von Anforderungen und Merkmalen, dargestellt mit Doppelpfeilen. Einzelpfeile zeigen eine entsprechende Verbesserung in Pfeilrichtung. Bespielweise verbessert sich die Kontrolle über die Hardware, je näher man in die Feldebene geht.

» **Cloud (public/private, off-premise):** Eine zentralisierte Cloud ermöglicht die gemeinsame Nutzung von Computing-Ressourcen durch verschiedene Mandanten (Endbenutzer, Anwender, Maschinen). Dadurch können Computing-Ressourcen durch mehrere Mandanten effizient miteinander geteilt werden. Diese Ressourcen befinden sich üblicherweise in räumlich weit entfernten Rechenzentren, welche von Cloud-Anbietern betrieben werden. Für den Zugriff wird in der Regel eine Internetanbindung benötigt, welche Stand heute keine Echtzeit-Kommunikation unterstützt, sondern nur Best-Effort-Service bietet.

» **Regional Edge Cloud (public/private, off-premise):** Stellt Rechenkapazität zur Verfügung, welche in der Regel regional, also räumlich deutlich näher zur jeweiligen Anwendung ist als eine zentralisierte Cloud. Regional Edge Clouds können an unterschiedlichen physischen Standorten miteinander und/oder mit einer zentralisierten Cloud verbunden sein. Die Regional Edge Cloud ist eine optionale Ebene (engl. *tier*), die weitaus geringere Latenzen zur Anwendung als die zentralisierte Cloud hat. Mittels dedizierter Verbindung kann zusätzlich weiche Echtzeitfähigkeit zur Verfügung gestellt werden.

» **Industrial Edge Cloud (private, on-prem):** Eine Industrial Edge Cloud verfügt über direkten oder indirekten Zugriff auf die Feldebene, was u. a. durch industriespezifische IT-Technologien wie Echtzeitbetriebssysteme (engl. *real-time operating system*) und *Time-Sensitive Networking* (TSN) unterstützt wird. Die Industrial Edge Cloud stellt eine skalierbare, unter Umständen ausfallsichere und verteilte Rechenleistung zur Verfügung, die von einem oder mehreren Edge Nodes bereitgestellt wird, um zeitkritische Aufgaben, welche u. U. Echtzeitanforderungen haben, durchzuführen. Der Ausfall einzelner Edge Nodes kann ggf. kompensiert werden, da mehrere Edge Nodes im Rechnerverbund arbeiten, um so die Ausfallsicherheit zu erhöhen. Die Funktionalität der Industrial Edge Cloud

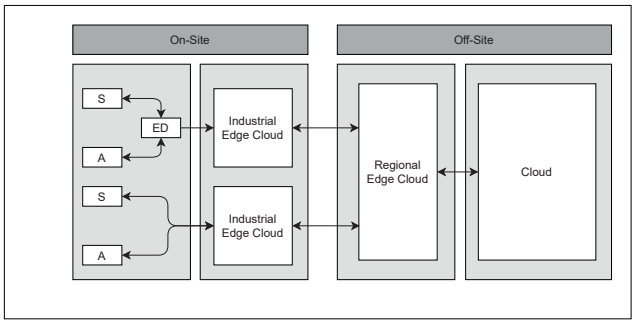

Abbildung 2: Mögliche Positionierung und Datenaustauschkanäle von Komponenten im Edge Computing Stack.

ist auch ohne Internetverbindung gewährleistet und kann mit der gleichen Cloud-Technologie aufgebaut werden. Im Unterschied zur Edge Cloud und zentralisierten Cloud, wird die Industrial Edge Cloud als Spezialfall der Private Cloud durch das jeweilige Unternehmen selbst betrieben.

» **Edge Node:** Ein Gerät, das Schnittstellen bereitstellt, um direkt oder indirekt (durch andere Edge Nodes) mit einem oder mehreren Feldgeräten verbunden zu werden und dadurch das Computing in dessen Nähe ermöglicht. Ein Edge Node ist ein computingfähiges Gerät und kann z. B. ein Ein-Platinen-Rechner, eine SPS, ein Industrie-PC, ein Server oder ein Computing Cluster sein. Dieser Node kann auch zur Filterung oder Vorverarbeitung der Daten verwendet werden, um diese später in die Cloud oder auf leistungsfähigere Computer zu übertragen. Je nach Anwendungsfall kann ein Edge Node eine Echtzeit- oder Nicht-Echtzeit-Verbindung mit den Geräten haben. Die Echtzeitfähigkeit kann zudem zwischen weicher Echtzeit und harter Echtzeit variieren. Ein Ausfall in harten Echtzeitanwendungen ist meist fatal, während in weichen Echtzeitanwendungen der Ausfall bis zu einem gewissen Punkt toleriert werden kann.

Industrielle Szenarien erfordern, je nach Anwendungsfall, Echtzeitfähigkeit und Datensicherheit. Edge Computing bzw. Industrial Edge Computing ist ein geeigneter Ansatz, um diesen Anforderungen gerecht zu werden. Um die Funktionalitäten des Industrial Edge Computing zu realisieren, sollte es durch eine Referenzarchitektur unterstützt werden, die im Folgenden beschrieben wird.

1.2 Security in der Edge

Eine digitalisierte und vernetzte Infrastruktur ist die Grundlage für einen optimalen Datenaustausch und eine effiziente Nutzung von Daten in der Produktion. Edge Nodes (EN) stehen als datenverarbeitende Systeme (wie auch Controller oder Cloudservices) als Teilnehmer in der Kommunikationskette zwischen den datenerzeugenden (Sensoren, S) und den durch Daten gesteuerten Devices (Aktoren, A). Im Zuge dieses Datenaustausches werden unterschiedliche Zonen über definierte Schnittstellen und Kanäle durchschritten (s. Abbildung 2). Begründet durch die Zugehörigkeit der Edge Nodes zu Maschinen und Anlagen sowie deren Steuerungen innerhalb der gleichen oder benachbarten Security Zone, besteht eine enge Verbindung zu Sensoren und Aktoren und damit die Möglichkeit der Einflussnahme auf das Maschinen- oder Anlagenverhalten und erfordert somit eine klare Beschreibung der Aufgaben und Grenzen dieser Technologien [6].
Ein Edge Node darf im Zusammenhang mit Cybersecurity nicht für sich isoliert, sondern muss im Gesamtkontext des Systems und sogar der ganzen Kommunikationskette betrachtet werden. Dies ist kongruent zu Cybersecurity-Konzepten wie *Defence in Depth* [6], das einen vielschichtigen Abwehrmechanismus darstellt. Um die Anforderungen an eine abgesicherte, vernetzte Umgebung ausreichend definieren zu können, müssen für jeden Teilnehmer, der am Datenaustausch beteiligt ist, die durch eine Cyber-Bedrohung entstehenden Risiken betrachtet und ihre möglichen Auswirkungen ermittelt werden. Eine Cyberattacke im Maschinenumfeld kann zu unterschiedlichen Auswirkungen führen [7]:

» Beeinträchtigung der Produktivität (Herbeiführen von Unproduktivzeiten)

» Diebstahl von geistigem Eigentum (Auslesen von bspw. Werkstückgeometriedaten oder Rezeptur eines Medikamentes)

» Produktsicherheit (Beeinträchtigung der Produktqualität oder fehlerhafte Verpackungszuordnung mit negativen Auswirkungen auf die Kunden)

» Maschinen- und Anlagensicherheit (Gefährdung von Mitarbeitern und Investitionsgütern)

Diese Auflistung zeigt, dass die gestiegene, systemische Komplexität einer vernetzten Produktionsumgebung die zu bewertenden Schutzziele im Rahmen einer holistischen Risikobeurteilung deutlich erweitert hat. Bisher rein Safety-fokussierte Risikobeurteilungen müssen mit dem Blick auf notwendige Security-Maßnahmen neu überdacht und gemeinsam betrachtet werden. Die Auswirkungen einer Manipulation auf eine Komponente und auf das ganze System müssen hinsichtlich ihrer Kritikalität bewertet und den am Anfang festgelegten Schutzzielen gegenübergestellt werden. Um die möglichen Angriffspunkte einer Cyberattacke identifizieren zu können, wird jede Komponente auf potenzielle Schwachstellen untersucht. Durch geeignete Gegenmaßnahmen [8] können die ermittelten Cyberrisiken auf ein akzeptables Maß reduziert werden.

2. Stand der Technik

Die Abstammung des Edge Computing reicht bis in die 1990er Jahre zurück, als Akamai Technologies *Content Delivery Networks* (CDN) einführte, um die Webleistung zu steigern [14]. Sie haben Inhalte am Rand des Internets zwischengespeichert, um Anfragen an die eigene Infrastruktur der Website zu reduzieren und den Benutzern schnellere Reaktionszeiten zu ermöglichen. Sie beantworteten ihre Anfragen mithilfe von Servern vor Ort.

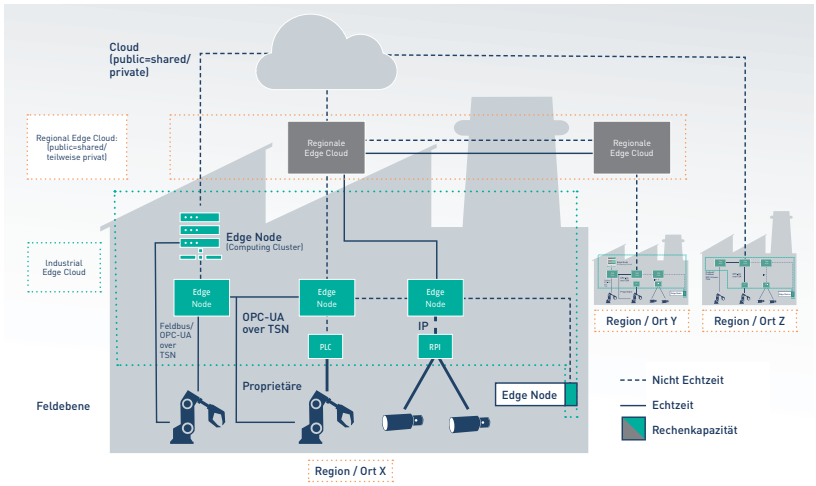

Abbildung 3: SmartFactory-KL Industrial Edge Cloud – Beispielarchitektur.

Noble et al. [9] demonstrierte erstmals das Potenzial von Edge Computing durch die Realisierung eines Spracherkennungsszenarios auf ressourcenbegrenzten Geräten. Sie haben die Berechnung auf einen nahegelegenen Server ausgelagert und die Ergebnisse lieferten eine angemessene Leistung.
Im Jahr 2012 führten Bonomi et al. [10] ein neues Paradigma namens *Fog Computing* ein. Im Gegensatz zu Edge Computing sind seine Teilnehmer ähnlich wie bei CDN über ein breiteres Netzwerk verteilt. Sie erklären auch die Notwendigkeit einer vereinheitlichenden Plattform, um eine verteilte Intelligenz zu schaffen.
Im Jahr 2014 haben Chang et al. [11] ein neues Modell für Cloud Computing mit dem Namen „Edge Cloud“ vorgeschlagen. Darin testeten sie die Leistung ihrer Architektur mit einer Indoor-Lokalisierungsanwendung, um die Latenz zu bewerten, und mit einer Videoüberwachungsanwendung, um die Bandbreite zu messen. Ihre Ergebnisse zeigten eine bessere Leistung im Vergleich zu den bestehenden Cloud-Lösungen.
Eines der größten Probleme von Edge Computing ist das Fehlen allgemein akzeptierter Standards [12].
In [9] wurden die bestehende Architekturen analysiert und industrielle Anforderungen berücksichtigt, um eine generische Software-Referenzarchitektur für die Industrial Edge Cloud zu bilden. Die Architektur ist dezentralisiert, herstellerunabhängig, kollaborativ, modular, erweiterbar, echtzeitfähig und unterstützt mehrere Nutzer. Ziel dieses Beitrags ist es, diese in der Praxis anhand von zwei Use Cases zu erproben.

3. Industrielle Edge-Cloud-Referenzachitektur

Basierend auf den vorgehenden Überlegungen kann eine Beispielarchitektur für die Edge Cloud Instanzen wie in Abbildung 3 erstellt werden. Eine Implementierung basierend auf dieser Beispielarchitektur wird in dem *Shared Production Use Case* in der SmartFactory-KL implementiert [14].
In Abbildung 3 sind einige der möglichen Kommunikationskombinationen von Cloud, Regional Cloud, Industrial Edge Cloud mit ihren Edge Nodes und vorhandenen Feldgeräten abgebildet. Die durchgezogenen Linien stellen die Echtzeitkommunikation dar, während die gestrichelten Linien die Nicht-Echtzeitkommunikation abbilden. Edge Nodes können über Echtzeit- oder Nicht-Echtzeit-Kommunikation miteinander kommunizieren. Sie können auch mit der regionalen Cloud oder nur mit der Cloud kommunizieren. Die Echtzeitkommunikation kann jede der verfügbaren Technologien sein, wie z. B. echtzeitfähige Feldbus, OPC-UA über TSN und/oder andere proprietäre Technologien. Ähnlicherweise kann jedes Übertragungsmedium verwendet werden, falls es echtzeitfähig ist. Jeder Edge Node kann unterschiedliche Hardware- und Softwarespezifikationen haben sowie von verschiedenen Anbietern bereitgestellt werden. Dies erfordert und ermöglicht Interoperabilität und verhindert auch Vendor-Lock-in-Probleme.
Die Industrial Edge Cloud erfordert auch Skalierbarkeit im Bereich der Integration von Edge Nodes. Dies kann durch die Orchestrierung des gesamten Netzwerks erreicht werden, indem bedarfsgerecht dezentrale Edge Nodes integriert oder desintegriert werden. Wenn eine Änderung erkannt wird, werden die Hardware- und Software-Spezifikationen unter allen Edge Nodes im Netzwerk geteilt. So kann jede Maßnahme auf jedem dezentralen Edge Node gleich sein.
Der Ort zum Ausführen der Anwendung oder Aufgabe hängt von der Latenzempfindlichkeit und dem Ressourcenbedarf ab. Wie aus Abbildung 1 ersichtlich, erfolgt die Ausführung im Falle einer Echtzeitanforderung auf Edge Cloud-Ebene. Wenn z. B. die Latenz weniger wichtig ist, kann die Anwendung in der Regional Edge Cloud und/oder der Cloud bereitgestellt und ausgeführt werden. Ein Orchestrator kann in diesem Fall per Standortentscheidung automatisch unterstützen, falls die Anforderungen vom Softwareentwickler festgelegt werden. Oder die Entscheidung vom Ausführungsort kann rein beim Softwareentwickler getroffen werden, wie z. B. in den Use Cases, die in Kapitel 5 erklärt werden.

3. Failover zwischen Industrial Edge Cloud und (Regional) Cloud

Die Umschaltung von einer gestörten auf eine betriebsbereite IT-Infrastruktur oder -Anwendung wird in Fachkreisen als *Failover* bezeichnet. Dieser Failover-Fall kann ein sehr komplexer Prozess werden. Deshalb ist dieser schon bei der

Entwicklung, Installation und Konfiguration einer IT-Infrastruktur oder bei der Software-Auswahl und -Installation zu beachten.
Das einfachste Szenario wäre der Ausfall einer Server-Infrastruktur und die sofortige Bereitstellung einer Ersatzserver-Infrastruktur. Diese Lösungsvariante wird meist mit der Eigenschaft *Cold-Standby* oder aktiv/passiv beschrieben. Cold-Standby-Modus bedeutet, dass der virtuelle Server, Container oder Betriebssysteme laufen aber die Anwendung nicht. Dadurch stehen die Anwendungen nicht ohne Unterbrechung bereit.
Für die unterbrechungsfreie Nutzung gibt es verschiedene Prinzipen. Die Lösung von Nutanix etwa besteht aus mindestens drei Servern für eine hyperkonvergente IT-Infrastruktur. Jeder Server ist mit zwei Netzwerk-Anschlüssen an zwei Netzwerk-Switches angeschlossen. Anwendungen werden über drei verschiedene Server verteilt. Sollte ein Server gestört sein, wird die Anwendungen ohne Unterbrechung auf einem weiteren Server aus dem gleichen Cluster zur Verfügung gestellt.
Egal welches Konzept für den Failover-Fall oder die Hochverfügbarkeit gewählt wird, es muss unbedingt sichergestellt werden, dass die notwendigen Daten an beiden Cloud-Standorten synchron zur Verfügung stehen. Dazu muss der Standort des Regional-Cloud-Rechenzentrums innerhalb eines bestimmten Radius gewählt werden. Die Angabe in Kilometer für den Radius bieten nur eine grobe Kennzahl. Viel wichtiger ist die reale Latenz zwischen den beiden Clouds, die auf dem Verbindungsweg zwangsläufig entsteht. Die Latenz (Zeitverzögerung) für ein Datenpaket, das in der Edge Cloud verarbeitet und gespeichert wird und das dann auch auf der Regional Cloud gespeichert wird, darf nicht größer als eine bestimmte Zeit sein. Dieser Wert liegt z. B. für Nutanix bei 5 ms.
Es gibt viele Anwendungsbeispiele und Möglichkeiten, die im Failover-Fall eine Umschaltung zwischen Industrial Edge Cloud und Regional Cloud notwendig machen oder auslösen können.
Eine technische Störung der IT-Infrastrukturen am Standort der Edge Cloud wäre ein Beispiel dafür, wie sinnvoll die Nutzung einer Regional Cloud sein kann. Gründe für Störungen könnten Teilausfall, Totalausfall, Wartungsarbeiten oder auch ein Hackerangriff sein.
Der Auslöser für einen Failover-Fall setzt ein messbares Ereignis und einen Ablaufplan voraus. Ein messbares Ereignis kann z. B. durch ein Monitor-System ermittelt werden, welches von außen regelmäßig eine Prüfung der CI oder Anwendung durchführt und bei festgelegten Abweichungen eine Aktion auslöst.
Einige Anwendungen verfügen über integrierte Lösungen für die Realisierung von Hochverfügbarkeit und den Failover-Fall. Ein Beispiel hierfür ist die Datenbankanwendung MySQL. Es ist auch möglich, zentralisierte oder dezentralisierte Orchestratoren zu haben, die sich mit der Entscheidung befassen, wo die Anwendung/Aufgabe ausgeführt werden soll.
Abhängig von der Netzwerk- und Softwarearchitektur und unabhängig davon, ob Edge Nodes direkt oder indirekt verbunden sind, können ihre Ressourcen auch gemeinsam genutzt werden. Dies ermöglicht das Orchestrieren der Edge Nodes über ein *Cloud Framework*, welches aus den einzelnen Teilnehmern ein selbstorganisierendes Cluster erzeugt. Deshalb können die Ausführung der Aufgaben auf den Edge Nodes gemeinsam geteilt (engl. *offloaded*) werden, um die Auslastung auszugleichen oder Ausfallzeiten zu vermeiden.
Wenn eine SPS, die einen Roboter steuert, ausfällt, müssen andere echtzeitfähige Edge Nodes die Steuerung übernehmen. Je nach Ressourcenverfügbarkeit, Echtzeitfähigkeit und Latenzzeit kann/können der/die am besten geeignete(n) Edge Node(s) ausgewählt werden und die Aufgaben ausführen. Im Beispiel können andere verfügbare echtzeitfähige Edge Nodes die Steuerung des Roboters übernehmen. Die Aufgaben, die weniger oder nicht latenzempfindlich sind, z. B. Bildverarbeitung oder Qualitätskontrolle mittels KI, können stattdessen auch in die Regional Edge Cloud oder in die Cloud (sofern vorhanden) übertragen oder *offloaded* werden.
Die Edge Nodes in der Architektur bieten und unterstützen die gängigen Schnittstellen (z. B. MQTT), um den Anwendungsbereich zu vergrößern. Wenn es mehrere Edge Nodes im Netzwerk gibt, erfolgt die Verwaltung als eigene Softwareebene zentral, einfach und sicher als separate Softwareschicht, wobei die Nodes weiterhin unabhängig voneinander dezentral arbeiten können.
Die Architektur schlägt folgende Schritte vor und befolgt diese, um den günstigsten Node für die Aufgaben-Ausführung zu wählen. Wenn ein Schritt keinen einzelnen Node bestimmen kann, bewertet der nächste Schritt die Situation. Die Schritte sind auch als Flussdiagramm in Abbildung 4 dargestellt.

1. Jeder Edge Node verfügt über eine aktualisierte Liste der verfügbaren Hardware-Ressourcen und Software zusammen mit ihren Standorten [4]. Zunächst fragt der aktuelle Node die möglichen Standorte für die angeforderte Aufgabe ab, einschließlich sich selbst.

2. Wenn der aktuelle Node die Software für diese Aufgabe nicht enthält oder nicht über genügend Ressourcen verfügt, dann wird ein alternativer Node im Netzwerk der Industrial Edge Cloud gesucht.

3. Falls mehrere Alternativen die Aufgabe rechtzeitig ausführen können, wird der Node gewählt, der die Aufgabe in der kürzesten Zeit ausführen kann.

4. Wenn die Ressourcen gleich sind, wird die Edge Node mit der geringsten Verzögerung gewählt.

5. Wenn auch die Latenzen vergleichbar sind, wird einer der Edge Nodes zufällig ausgewählt.

6. Wenn kein Edge Node gefunden wird, der die Aufgabe ausführen kann, dann versucht der aktuelle Node, die Ausführung zu planen, wobei „Nein“ in mehreren Alternativzweigen folgt. In diesem Fall wird die Ausführung in dem Edge Node durchgeführt, der die ursprüngliche Aufgabenanfrage erhalten hat. Planung bedeutet, dass der Edge Node z. B. die Ausführung der Aufgaben sortiert (engl. *scheduling*).

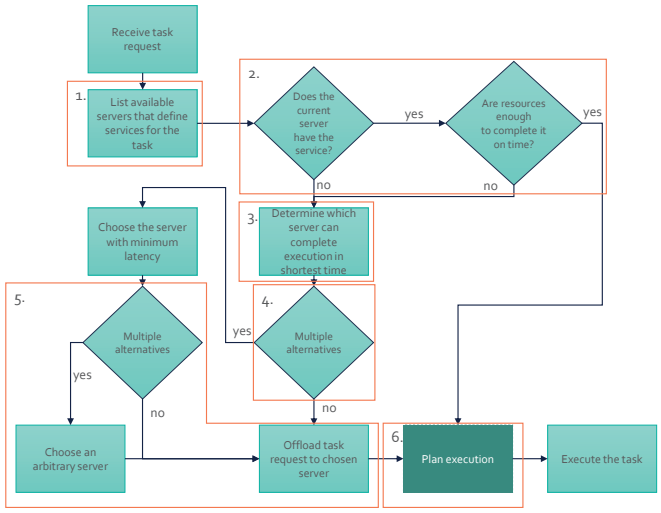

Abbildung 4: Dezentralisiertes Offloading Flussdiagramm zwischen Edge Nodes.

Falls es eine Regional Edge Cloud- oder Cloud-Verbindung gibt, evaluiert der Node auch, ob die Ausführung der Aufgabe an diesen Orten in der geforderten Zeit machbar ist. Ist dies der Fall, wird er auf den am besten geeigneten übertragen.

5. Use Cases in der SmartFactory-KL

5.1 Use Case: Qualitätskontrolle in der Edge und Cloud

Der *SmartFactory-KL Production Level 4-Demonstrator* (PL4-Demonstrator) ist ein komplexes Zusammenspiel aus etlichen IT- (ohne Echtzeitanforderungen) und OT-Services (mit Echtzeitanforderungen, sowohl harte als auch weiche). Derzeit sind IT-Services sowohl in der Industrial Edge Cloud als auch teilweise in der Public Cloud angesiedet (s. Abbildung 5). Hingegen sind OT-Services als abstrakte Fähigkeiten (bspw. *CheckQuality-Skill*) gekapselt, derzeit fest mit dem physikalischen Modul (bspw. Qualtitätskontroll-Modul) gekoppelt und befinden sich in der Feldebene. Genauer betrachtet ist dieser Skill jedoch zusammengesetzt (*Composite Skill*) aus mehreren atomaren Fertigkeiten (*Atomic Skills*), die nicht alle hardware-gebunden sein müssen [9].
Im angeführten Beispiel muss nur die Aufnahme des Qualitätskontrollbilds mit der Kamera des Moduls durchgeführt werden. Die anschließende Prüfung auf Beschädigung oder fehlende Bauteile erfolgt an anderer Stelle. Diese Trennung ermöglicht mehrere Verbesserungen:
Zum einen kann der atomare Überprüfungsskill auf mehreren (und leistungsstärkeren) Edge Nodes oder Regional Edge Clouds zur Verfügung appliziert werden, um Redundanz für den Fehlerfall zu schaffen und Produktionsstillstände zu vermeiden. Aufgrund der weichen Echtzeitanforderung (Zeitspanne des Transports zum nächsten Modul, ca. 5 s) kann diese jedoch nicht in beliebig höheren Ebenen geschehen, ohne Einfluss auf die Produktionsgeschwindigkeit zu haben. Zum anderen schränkt die vorhandene Rechenleistung der SPS im Modul die Komplexität und damit die Güte der Überprüfung ein. Leistungsstärkere Edge Nodes oder Regional Edge Clouds ermöglichen dahingehend eine deutliche Verbesserung, sofern die weichen Echtzeitanforderungen dennoch eingehalten werden können.

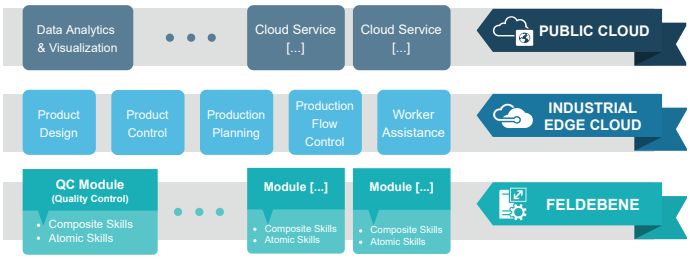

Abbildung 5: Komponentenansicht des Production Level 4 (PL4) Demonstrators [1].

Weiterhin erlaubt eine Cloud-basierte Lösung sowie eine *Shared Production* (s. Kapitel 5.2) Elastizität. In Zeiten hohen Bedarfs (und damit schneller Produktionsabfolgen) werden mehr Ressourcen benötigt als in Zeiten mit niedrigem Bedarf. Wenn traditionelle Lösungen nur in der Feldebene für die höchste Produktionsabfolge ausgelegt sind, werden diese bei niedrigem Bedarf nicht sinnvoll genutzt. Sind sie hingegen für eine mittlere Produktionsgeschwindigkeit dimensioniert, so limitieren diese im Fall erhöhten Bedarfs. Im obigen Fall ermöglicht die Skalierbarkeit eine kostenoptimierte Produktion. Durch die in Kapitel 1.2 beschriebenen Cybersecurity-Konzepte sind firmeninterne Daten dennoch vor fremden Zugriff geschützt.

5.2 Use Case: Shared Production

Die Vision der zukünftigen industriellen Produktion ist eine geteilte Produktion (Shared Production). Für jeden Auftrag ergeben sich neue Wertschöpfungsnetzwerke, bestehend aus unterschiedlichen Unternehmen, die sich zur Laufzeit zusammensetzen können. Die Konfiguration dieser Wertschöpfungsnetzwerke kann dabei über entsprechende Plattformen orchestriert werden. Die Vernetzung der einzelnen Unternehmen soll dabei in Zukunft über das Datenökosystem Gaia-X erfolgen.
Eine Besonderheit der Shared Production ist die Vernetzung des Unternehmens bis auf Maschinenebene. Dies tiefe Integration und Interaktion verschiedener Ökosystemteilnehmer erfordert zusätzliche Sicherheitsanforderungen an das genutzte IT-System. Zusätzlich zur Absicherung des Maschinenparks müssen auch Ausfälle in der IT-Infrastruktur, wie in den vorigen Kapiteln aufgezeigt, verhindert werden.
Für die Umsetzung dieser Vision und das Aufzeigen der möglichen Konzepte nutzt die SmartFactory-KL ihr eigenes Produktionsnetzwerk. Diese besteht aus vier Produktionsinseln, die jeweils ein Unternehmen mit spezifischem Serviceangebot darstellen. In der modularen Produktionsumgebung wird ein Modell-LKW aus Noppensteinen gefertigt. Dieser besteht aus vier Baugruppen, welche gesondert auf den jeweiligen Produktionsinseln gefertigt werden und somit das Prinzip der Shared Production aufzeigen. Wie in der Realität wird der LKW individuell konfiguriert. Dadurch wird eine Losgröße-Eins-Produktion erreicht, welche eine dynamische und modulare Produktionsumgebung benötigt. Im Fokus der Entwicklung stehen dabei *Production-as-a-Service* (PaaS) und *AI-as-a-Service* (AIaaS). Diese Services werden für interne sowie externe Nutzer bereitgestellt.
In Abbildung 6 ist der konzeptionelle Entwurf des Systems zu sehen. Jede Produktionsinsel besteht dabei aus seiner

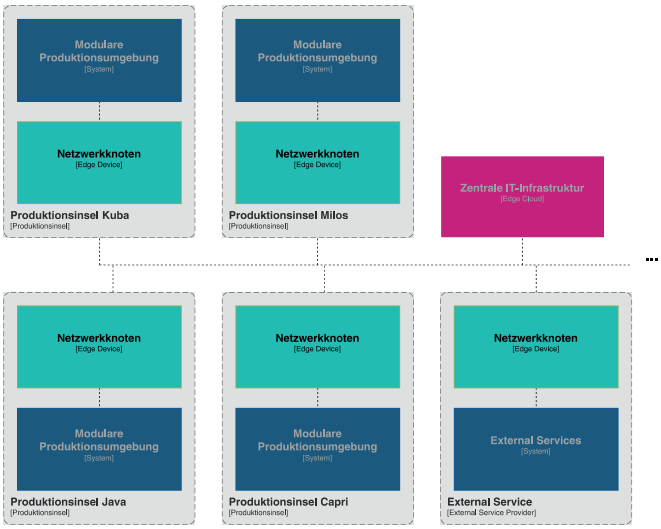

Abbildung 6: Produktionsnetzwerk der SmartFactory-KL mit seinen vier Produktionsinseln (Kuba, Milos, Java, Capri).

modularen Produktionsumgebung und einem Netzwerkknoten, welcher durch Edge Nodes realisiert wird. Das Edge Node stellt die zentrale Kommunikationsschnittstelle der Produktionsinsel zum Netzwerk dar. Der Austausch der Informationen erfolgt hierbei über den *International Data Space Connector* (IDS), der einen Datenraum für die Produktionsumgebung aufbaut. Zusätzlich zum reinen Datenaustausch laufen auch containerisierte Services, wie Produktions-Services, KI-gestützte Qualitätssicherung (s. Kapitel 5.1) oder Datenanalyse-Services auf dem Edge-Device, welche zum Betrieb der Produktionsinsel notwendig sind.
Bestimmte Services sind aufgrund ihrer Anforderungen an die Produktinsel gebunden und müssen ausfallsicher betrieben werden. Dagegen können Datenservices und KI-Services auch auf der Edge Cloud oder Cloud betrieben werden. Grundsätzlich ist ein Betrieb der Services auf der Produktionsinsel selbst sinnvoll, da hier die kürzesten Latenzzeiten erreicht werden und somit Anforderungen an weiche und harte Echtzeit erfüllt werden können. Kommt es zum Fehlerfall, ist eine Failover-Strategie notwendig. Hier können containerisierte Softwarelösungen schnell auf eine übergeordnete Instanz transferiert und eine Nutzung dessen sichergestellt werden.

6. Zusammenfassung und Ausblick

In diesem Beitrag wurde erläutert, wie die Industial Edge Cloud definiert ist, einschließlich ihrer Komponenten. Zudem wurde anhand eines Anwendungsbeispiels gezeigt, wie diese für die Industrie ausgestaltet werden kann. Abhängig von den Zeit- und Ressourcenanforderungen der Aufgabe, z. B. wenn sie innerhalb einer bestimmten Zeitspanne ausgeführt werden muss, bestimmen Edge Nodes selbstständig den optimalen Ort für die Ausführung. Falls ein Edge Node nicht über genügend Ressourcen zur Ausführung der Aufgabe verfügt, wird ein anderer Edge Node im gleichen Netzwerk gesucht, der diese Aufgabe rechtzeitig ausführen könnte. Um den optimalen Edge Node zu bestimmen, werden die Hardware-Ressourcen, die Verbindungstechnologien (IP, 5G, Glasfaser, etc.) und die Echtzeitfähigkeiten ausgewertet. Die Aufgabenausführung wird dann an die ausgewählten Nodes weitergeleitet. Alternativ wäre die Weiterleitung der Aufgabe an eine höhere Ebene, z. B. eine regionale Edge Cloud oder eine Cloud, ebenfalls möglich, wenn diese verfügbar ist. Wie erwartet, wird bei einer Echtzeitanforderung an eine Aufgabe die Nutzung der Cloud aus technischen Gründen vermieden.
Um die Interoperabilität zu erhöhen und Vendor-Lock-in-Probleme zu vermeiden, wird ein offenes Cloud Framework für den zuvor definierten Anwendungsfall implementiert werden. Das Ziel des Anwendungsfalls ist es, die Produktion mit einer höheren Rate am Laufen zu halten, falls der lokale Qualitätskontrolldienst ausfällt.
Als zukünftige Arbeit ist geplant, die vorgeschlagene Architektur in den Anwendungsfall zu integrieren und die Leistung der verschiedenen Ebenen in verschiedenen Ausfallszenarien zu bewerten.

Danksagung

Diese Forschung wurde teilweise durch das H2020-Programm der Europäischen Union unter der Projektnummer 101017047 (Projekt PHYSICS) finanziert. Die Verantwortung für diese Veröffentlichung liegt bei den Autoren.

Referenzen

[1] Satyanarayanan, M. (2017). The emergence of edge computing. *Computer, 50*(1), 30-39.
[2] Bai, H. (2014). *Zen of Cloud*. Florida, Routledge.
[3] Amazon Web Services, Inc. (2022). *Weltweite Infrastruktur*. Abgerufen von: https://aws.amazon.com/about-aws/global-infrastructure/?nc1=h_ls.
[4] Gezer, V., Um, J., Ruskowski, M. (2018). An introduction to edge computing and a real-time capable server architecture. *Int. J. Adv. Intell. Syst.(IARIA), 11*(7), 105-114.
[5] Linux Foundation Edge. (2020). *Sharpening the Edge: Overview of the LF Edge Taxonomy and Framework*. Abgerufen von: https://www.lfedge.org/wp-content/uploads/2020/07/LFedge_Whitepaper.pdf
[6] in IEC/TS 62443-1-1. (2009). Industrial communication networks - Network and system security - Part 1-1: Terminology, concepts and models. IEC: www.iec.ch
[7] Bundesamts für Sicherheit in der Informationstechnik (BSI). (2021). *Die Lage der IT-Sicherheit in Deutschland 2021*. Abgerufen von: https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/bsi-lagebericht-cybersicherheit-2021.pdf?__blob=publicationFile&v=3.
[8] IEC 62443-3-3. (2020). Industrielle Kommunikationsnetze – IT-Sicherheit für Netze und Systeme. IEC: www.iec.ch
[9] Ruskowski, M., Herget, A., Hermann, J., Motsch, W., Pahlevannejad, P., Sidorenko, A., ... Wagner, A. (2020). Production Bots für Production Level 4: Skillbasierte Systeme für die Produktion der Zukunft. *atp magazin, 62*(9), 62-71.
[10] Siwach, V. (2020). *TheNewStack*. Abgerufen von: https://thenewstack.io/sharpening-the-edge-the-linux-foundation-edge-framework-and-taxonomy/.
[11] Eagar, M. (2020). What is the difference between decentralized and distributed systems?. Abgerufen von: https://medium.com/distributed-economy/what-is-the-difference-between-decentralized-and-distributed-systems-f4190a5c6462.

[12] Kendrick, B. A., Dhokia, V., Newman, S. T. (2017). Strategies to realize decentralized manufacture through hybrid manufacturing platforms. *Robotics and computer-integrated manufacturing, 43*, 68-78.

[13] Wagner, T., Hausner, C., Elger, J., Lowen, U., Luder, A. (2010). *Engineering processes for decentralized factory automation systems*. IntechOpen.

[14] Dilley, J., Maggs, B., Parikh, J., Prokop, H., Sitaraman, R., Weihl, B. (2002). Globally distributed content delivery. *IEEE Internet Computing, 6*(5), 50-58. https://doi.org/10.1184/R1/6605972.v1

AUTOREN

Volkan Gezer, Dr.-Ing. (geb. 1989) ist seit 2015 wissenschaftlicher Mitarbeiter am Deutschen Forschungszentrum für Künstliche Intelligenz. Seine aktuellen Forschungsschwerpunkte sind Edge Computing, Cloud Computing und dezentralisierte Orchestrierung.

Dr.-Ing. Volkan Gezer
Deutsches Forschungszentrum für Künstliche Intelligenz
Trippstadter Str. 122
67663 Kaiserslautern
✆ +49 631 20575 1063
@ volkan.gezer@dfki.de

Carsten Harms, M.Sc. (geb. 1989) ist seit 2017 wissenschaftlicher Mitarbeiter bei der Technologie-Initiative SmartFactory-KL e.V. und seit 2019 am Deutschen Forschungszentrum für Künstliche Intelligenz. Sein aktueller Forschungsschwerpunkt ist KI-basierte Zustandsüberwachung/Predictive Maintenance unter Nutzung von Edge- und Cloud-Computing.

Carsten Brüggemann (geb. 1967) ist seit 2000 Mitarbeiter der PFALZKOM GmbH. Seine Arbeitsschwerpunkte sind praktische Lösungsentwicklung für Managed Services im Cloud Umfeld. Dazu gehören auch Netzwerk-, Sicherheits-, I-IoT und Hyperkonvergente Infrastruktur Technologien.

Carsten Brüggemann
PFALZKOM GmbH
Koschatplatz 1
67061 Ludwigshafen
✆ +49-621 5853173
@ carsten.brueggemann@pfalzkom.de

Michael Pfeifer (geb. 1984) ist Maschinenbauingenieur und arbeitet seit 2009 bei TÜV SÜD. Sein Arbeitsschwerpunkt ist die Sicherheit von Maschinen entlang des Lebenszyklusses. Im Rahmen der Weiterentwicklung der Maschinensicherheit, um den Ansprüchen von I4.0 gerecht zu werden, beschäftigt er sich mit dem Zusammenführen von Safety und Cybersecurity und als „TÜV SÜD Smart Safety Lead Architect“ mit der Dynamisierung der Safety.

Michael Pfeifer
TÜV SÜD Industrie Service GmbH
Westendstraße 199
80686 München
✆ +49 89 5791-3329
@ michael.pfeifer@tuvsud.com

Andreas Michael (geb. 1969), Ingenieur für Elektrotechnik, arbeitet als Industrial IT Security Expert beim TÜV SÜD. Seine beruflichen Stationen im Maschinenbau und der IT, führten ihn früh zum Thema Industrie 4.0, welches er seitdem in Wirtschaft und Wissenschaft mitgestalten durfte. Er arbeitet an der Integration von Security in die Industrie, vom Shopfloor bis in die Cloud und unterstützt bei der Gremienarbeit zur normativen Umsetzung des Themas.

Simon Althoff (geb. 1982) ist seit 2017 Mitarbeiter bei der Weidmüller Interface GmbH & Co. KG. Als Technologieentwickler beschäftigt er sich unter anderem mit der Konzeptionierung und dem Betrieb von skalierbaren Edge Clouds im industriellen Umfeld.

Simon Althoff
Weidmüller Interface GmbH & Co. KG
Klingenbergstraße 26
32758 Detmold
✆ +49 5231 14 293249
@ simon.althoff@weidmueller.com

Dr. Torsten Runge (geb. XXXX) ist Senior Architect bei der Deutschen Telekom, T-Systems. Seine Arbeitsschwerpunkte fokusieren sich auf Kommunikationsnetze und -systeme, insbesondere in den Bereichen Cloud und Edge Computing, Software-Defined Networks (SDN) und Network Function Virtualization (NFV) mit Anwendungen für Interntet of Things (IoT), Cyber-Physical Systems (CPS) und Autonomous Systems.

Prof. Dr.-Ing. Martin Ruskowski (geb. 1969) ist Forschungsbereichsleiter am Deutschen Forschungszentrum für Künstliche Intelligenz, Inhaber des Lehrstuhls für Werkzeugmaschinen und Steuerungen an der TU Kaiserslautern und Vorstandsvorsitzender der im DFKI beheimateten Technologie-Initiative SmartFactory KL e.V. Sein Forschungsschwerpunkte sind neuartige Steuerungskonzepte für die Automatisierung und der Einsatz Künstlicher Intelligenz in der Automatisierungstechnik.

Keran Sivalingam, M.Sc. (geb. 1988) ist seit 2019 wissenschaftlicher Mitarbeiter bei der Technologie-Initiative SmartFactory-KL e.V. Seine aktuellen Forschungschwerpunkte sind Production-as-a-Service und Produktionsplanung im Kontext der Shared Production.