Lokalkammer München
UPC_CFI_180/2025
UPC_CFI_210/2025

Entscheidung

des Gerichts erster Instanz des Einheitlichen Patentgerichts
Lokalkammer München
verkündet am 11. März 2026

LEITSÄTZE

Stellungnahmen eines Patentanmelders im Erteilungsverfahren stellen kein Auslegungsmaterial für die Auslegung des erteilten Patents im Sinne von Art. 69 Abs. 1 EPÜ dar. Ebenso wenig sind solche Äußerungen des Anmelders im Erteilungsverfahren für die Auslegung in einem nachfolgenden Verletzungs- oder Nichtigkeitsverfahren bindend. Sie können jedoch ein Indiz für das fachmännische Verständnis von der Lehre des erteilten Patents im Prioritätszeitpunkt geben, weil der Anmelder selbst regelmäßig das beste Verständnis von seiner Erfindung und das zugehörige Fachwissen hat (Fortführung von: Berufungsgericht, Anordnung vom 20.12.2024, UPC_CoA_402/2024 - Alexion gg. Samsung Bioepis).

Es stellt keine Klageänderung im Sinne von Regel 263 VerfO dar, wenn sich das angegriffene Produkt nach der Klageerwiderung des Beklagten in technischen Einzelheiten anders darstellt als es der Kläger in der Klageschrift beschrieben hatte und der Kläger den Verletzungsvorwurf in der Replik auf die vom Beklagten beschriebene technische Funktionsweise stützt, solange sich der Verletzungsangriff weiterhin auf das allgemein beschriebene Produkt und seine angegriffene Funktion bezieht.

Es ist prozessual zulässig, dass der Beklagte einer Verletzungsklage eine Widerklage auf Nichtigerklärung unter der auflösenden Bedingung erhebt, dass die Verletzungsklage keinen Erfolg hat. Tritt diese Bedingung ein, weil das Streitpatent ungeachtet des Rechtsbestands des Streitpatents nicht verletzt ist, ist über die Widerklage auf Nichtigerklärung nicht mehr zu entscheiden.

HEADNOTES

Statements made by a patent applicant during the grant procedure do not constitute material for the interpretation of the granted patent within the meaning of Art 69(1) EPC, nor are such statements by the applicant in the grant procedure binding for the claim construction in subsequent infringement or revocation proceedings. However, they may provide an indication of how a person skilled in the art would understand the technical teaching of the granted patent on the priority date, since the applicant himself usually has the best understanding of his invention and the associated technical knowledge (continuation of CoA, Order of 20 December 2024, 402/2024 - Alexion v Samsung Bioepis).

It does not constitute an amendment to the claim within the meaning of Rule 263 of the RoP if, according to the Defendant's Statement of Defence, the attacked product differs in technical details from that described by the Claimant in the Statement of Claim and the Claimant bases the allegation of infringement in the Reply to the Statement of Defence on the technical functioning described by the Defendant, as long as the infringement claim continues to relate to the generally described product and its attacked function.

It is admissible in procedural terms for the Defendant to file a counterclaim for revocation in response to an infringement action, subject to the condition subsequent that the infringement action is unsuccessful. If this condition is met because the patent at issue is not infringed regardless of its validity, no decision is required on the counterclaim for revocation.

KLÄGERINNEN, BEKLAGTE ZU 2) ZUGLEICH WIDERBEKLAGTE

1. **BFexaQC AG**, vertreten durch ihren Vorstandsvorsitzenden Bernhard Frohwitter, Südliche Münchner Straße 56, 82031 Grünwald, Deutschland,

2. **ParTec AG**, vertreten durch ihren Vorstandsvorsitzenden Bernhard Frohwitter, Possartstraße 20, 81679 München, Deutschland,

vertreten durch: Rechtsanwalt Dr. Roman Sedlmaier und Patentanwalt Jan Gigerich, IPCGS Gigerich Sedlmaier Patentanwalt Rechtsanwalt PartGmbB, Paul-Wassermann-Straße 3, 81829 München,

unterstützt durch: Dr. Stefan Richter, Clifford Chance Partnerschaft mbB, Königsallee 59, 40215 Düsseldorf,

Patentanwälte David Molnia und Alexandre Hoffmann, df-mp tech Molnia Ho PartG mbB, Theatinerstraße 16, 80333 München und

Rajvinder Jagdev und Peter FitzPatrick, Irish Solicitors, Powell Gilbert (Europe) LLP, 28-32 Pembroke Street Upper, Dublin, D02 EK84.

BEKLAGTE UND WIDERKLÄGERINNEN

1. **NVIDIA Corporation**, vertreten durch den Gründer, Präsidenten und CEO, Herrn Jensen Huang, 2788 San Tomas Expressway, Santa Clara, CA 95051, USA,

2. **NVIDIA GmbH**, vertreten durch jeden der Geschäftsführer, Mark Steven Hoose, Ludwig von Reiche, Rebecca Peters, Donald Robertson, Adenauer Straße 20/A4, 52146 Würselen, Deutschland,

vertreten durch: Johannes Heselberger, Dr. Christof Karl, Dr. Stefan Lieck, Dr. Philipp Bovenkamp, Sabrina Hütt und Henri Kirner, Rechtsanwälte und Patentanwälte der Bardehle Pagenberg Partnerschaft mbB, Prinzregentenplatz 7, 81675 München.

STREITPATENT

Europäisches Patent Nr. EP 3 743 812

SPRUCHKÖRPER/KAMMER

Spruchkörper 2 der Lokalkammer München

MITWIRKENDE RICHTER/INNEN

Diese Entscheidung wurde erlassen unter Mitwirkung des Vorsitzenden Richters Dr. D. Voß (Berichterstatter), des rechtlich qualifizierten Richters Dr. G. Werner, des rechtlich qualifizierten Richters A. Kupecz und des technisch qualifizierten Richters A. Dumont.

VERFAHRENSSPRACHE

Deutsch

GEGENSTAND

Verletzungsklage und Widerklage auf Nichtigerklärung

MÜNDLICHE VERHANDLUNG

13. Februar 2026

SACHVERHALT

1 Die Klägerinnen nehmen die Beklagten wegen Verletzung des Europäischen Patents mit einheitlicher Wirkung EP 3 743 812 (UPC_CFI_180/2025) in Anspruch. Die Beklagten haben gegen die Klägerin zu 2) Widerklage auf Nichtigerklärung des Streitpatents erhoben (UPC_CFI_210/2025).

2 Das Streitpatent wurde am 23. Januar 2019 von der Klägerin zu 2) unter Inanspruchnahme der Priorität der EP 18152903 vom 23. Januar 2018 angemeldet. Die Anmeldung wurde am 2. Dezember 2020 offengelegt. Der Hinweis auf die Erteilung des Streitpatents wurde am 2. August 2023 veröffentlicht. Die Registrierung als Patent mit einheitlicher Wirkung, beantragt am 30. August 2023, erfolgte am 6. September 2023 (vgl. Anlage K 45). Die Vertragsmitgliedsstaaten zum Zeitpunkt der Registrierung der einheitlichen Wirkung waren Österreich, Belgien, Bulgarien, Deutschland, Dänemark, Estland, Finnland, Frankreich, Italien, Litauen, Luxemburg, Lettland, Malta, Niederlande, Portugal, Schweden und Slowenien.

3 Das Streitpatent, dessen Verfahrenssprache Englisch ist, betrifft die über die Anwendungslaufzeit bestimmte dynamische Zuweisung von heterogenen Rechenressourcen. Anspruch 1 des Streitpatents schützt ein Verfahren und lautet wie folgt:

> „A method of operating a heterogeneous computing system (10) comprising a plurality of computation nodes (20) and a plurality of booster nodes (22), at least one of the plurality of computation nodes (20) and plurality of booster nodes (22) being arranged to compute a computation task, the computation task comprising a plurality of sub-tasks, wherein

in a first computing iteration, the plurality of subtasks are assigned to and processed by ones of the plurality of computation nodes (20) and booster nodes (22) in a first distribution;

characterized in that

information relating to the processing of the plurality of sub-tasks by the plurality of computation nodes (20) and booster nodes (22) is used to generate a further distribution of the sub-tasks between the computation nodes (20) and booster nodes (22) for processing thereby in a further computing iteration.

4 In der deutschen Übersetzung lautet Anspruch 1 des Streitpatents:

„Verfahren zum Betreiben eines heterogenen Rechensystems (10), das eine Vielzahl von Rechenknoten (20) und eine Vielzahl von Booster-Knoten (22) aufweist, wobei mindestens einer der Vielzahl von Rechenknoten (20) und der Vielzahl von Booster-Knoten (22) so eingerichtet ist, dass er eine Rechenaufgabe berechnet, wobei die Rechenaufgabe eine Vielzahl von Unteraufgaben aufweist,

wobei in einer ersten Iteration der Berechnung die Vielzahl von Unteraufgaben in einer ersten Verteilung einem der Vielzahl von Rechenknoten (20) und Booster-Knoten (22) zugewiesen und von diesen verarbeitet werden;

dadurch gekennzeichnet, dass

Informationen, die sich auf die Verarbeitung der Vielzahl von Unteraufgaben durch die Vielzahl von Rechenknoten (20) und Booster-Knoten (22) beziehen, verwendet werden, um eine weitere Verteilung der Unteraufgaben zwischen den Rechenknoten (20) und Booster-Knoten (22) zu erzeugen, um sie in einer weiteren Recheniteration zu verarbeiten.“

5 Die Klägerin zu 2) ist alleinige, im Patentregister eingetragene Inhaberin des Streitpatents.

6 Am 29. August 2021 unterzeichnete Herr Bernhard Frohwitter, Vorstandsvorsitzender beider Klägerinnen, für diese einen Vertrag, mit dem der Klägerin zu 1) eine ausschließliche Lizenz am Streitpatent für den Bereich der Mikroelektronik einschließlich der Entwicklung und Herstellung und des Vertriebs von Mikrochips und -prozessoren eingeräumt werden sollte. Wegen der Einzelheiten wird auf die Anlage K 50 Bezug genommen.

7 Ein weiterer Vertrag, mit dem der FL Systems AG & Co. KG eine ausschließliche Lizenz unter anderem am Streitpatent für den Anwendungsbereich der Systemarchitektur von Supercomputern (High Performance Computing) erteilt werden sollte, wurde am 21. Dezember 2023 von Herrn Frohwitter für die Klägerin zu 2) und die FL Systems AG & Co. KG unterzeichnet. Die Klägerin zu 2) ist Komplementärin der FL Systems AG & Co. KG. Herr Frohwitter unterzeichnete zudem am 7. Juni 2024 für die FL Systems AG & Co. KG und die

Klägerin zu 1) einen Vertrag, mit dem erstere die ihr erteilte ausschließliche Lizenz weiterlizenzierte. Wegen der Einzelheiten der beiden Lizenzverträge wird auf die Anlagen K 48 und K 49 verwiesen.

8 Die Beklagte zu 1) entwickelt Grafikprozessoren. Sie stellt Nvidia DGX-Systeme her, die sie unter anderem auf ihrer deutschsprachigen Website "https://www.nvidia.com/de-de/datacenter/dgx-systems/" unter Verweis auf die Produkte "Nvidia DGX A100", " Nvidia DGX H100", " Nvidia DGX BasePOD" oder " Nvidia DGX SuperPOD" anpreist (Anlage K 8). Im Rahmen der Darstellung ihrer DGX-Systeme weist die Beklagte zu 1) darauf hin, dass die Nvidia DGX Systeme über zertifizierte Partner erworben werden können, darunter für die Bundesrepublik Deutschland die „NPN Elite Lösungsanbieter“ „Amber“, „Delta“ und „sysGen“ (Anlagen K 9 und K 10). Mehrere hundert Lösungsanbieter aus einer Vielzahl von Staaten finden sich in der vollständigen Liste der NPN-Partner (Anlage K 11). In einem Blogbeitrag vom 1. Mai 2023 auf der Website der Beklagten zu 1) wird zudem der weltweite Einsatz der DGX H100-Systeme angepriesen, unter anderem für DeepL in der Bundesrepublik Deutschland (Anlage K12).

9 Die Beklagte zu 2) gehört zum Nvidia-Konzern. Sie wird von der Nvidia Limited mit Sitz in London kontrolliert, die wiederum in den Konzernabschluss der Beklagten zu 1) einbezogen ist. Gegenstand der Tätigkeit der Beklagten zu 2) ist laut Handelsregisterauszug die Entwicklung von Halbleitern und anderen Computerhardware- und Softwarekomponenten und der Vertrieb eigener und fremder Hard- und Softwareprodukte (Anlage K 6). Die Beklagte zu 2) ist auch in den Bereichen Forschung und Entwicklung sowie Verkaufsförderung und Marketing tätig (Anlage K 4). Sie übernimmt zudem die Gewährleistung von Produktzuverlässigkeit und Qualitätsmanagement für Hersteller, Montagefirmen und Distributoren, die die Nvidia-Technologie in ihren Produkten verwenden oder Nvidia-Produkte veräußern. Das Modell DGX H100 trägt zudem die CE-Kennzeichnung, an deren Erhalt die Beklagte zu 2) aktiv mitwirkt. Bei ihr als einziger europäischer Firma kann die Kopie der Konformitätserklärung direkt angefordert werden.

10 Die Klägerinnen wenden sich mit ihrer Klage gegen DGX-System-Produkte der Beklagten.

11 Nvidia DGX ist eine Reihe von Nvidia-Servern und -Workstations, die neben der Verwendung von CPUs schwerpunktmäßig auf die Verwendung von Grafikprozessoren (GPUs) zur Beschleunigung von Deep-Learning-Anwendungen spezialisiert sind. Das DGX-System "DGX H100" bildet den grundlegenden Baustein für große KI-Cluster wie DGX POD oder DGX SuperPOD und ist damit als Herzstück eines KI-Kompetenzzentrums für Unternehmen konzipiert. Andere DGX-Systeme wie DGX A100, DGX H200 oder DGX GH200 verwenden unterschiedliche Komponenten, unterscheiden sich aber konzeptionell nicht vom DGX H100 Modell. Sämtliche DGX-Server und -Workstations sind mit Blick auf den Verletzungsvorwurf identisch. Sie enthalten GPUs in Kombination mit CPUs mit einer Vielzahl von CPU-Cores. Wenn diesen Produkten Arbeitslasten zugewiesen werden, fungieren die CPUs als zentrale

Verwaltungsstelle für die GPUs, verteilen die Aufgaben und sammeln die verfügbaren Ergebnisse.

12 Auf den Nvidia DGX Systemen kann die Run:ai-Software des Unternehmens Run:ai installiert werden. Die Run:ai Cluster Engine nimmt die dynamische Verwaltung von KI-Arbeitslasten und -Ressourcen insbesondere für optimale GPU-Auslastung vor. Sie setzt sich aus zwei Komponenten zusammen: einerseits die Run:ai-Cluster für Scheduling-Dienste und das Workload-Management und andererseits die Run:ai-Kontrollebene für das Ressourcenmanagement, die Workload-Übermittlung und die Cluster-Überwachung. Zum Jahreswechsel 2024/2025 übernahm der Nvidia-Konzern das Unternehmen Run:ai.

ANTRÄGE

13 Mit der Verletzungsklage beantragen die Klägerinnen:

I. Die Beklagten werden verurteilt, es zu unterlassen,

Vorrichtungen, die zur Durchführung eines Verfahrens zum Betreiben eines heterogenen Rechensystems (10) geeignet sind, das eine Vielzahl von Rechenknoten (20) und eine Vielzahl von Booster-Knoten (22) aufweist, wobei mindestens einer der Vielzahl von Rechenknoten (20) und der Vielzahl von Booster-Knoten (22) so eingerichtet ist, dass er eine Rechenaufgabe berechnet, wobei die Rechenaufgabe eine Vielzahl von Unteraufgaben aufweist,

Abnehmern im Hoheitsgebiet der Republik Österreich, des Königreichs Belgien, der Republik Bulgarien, des Königreichs Dänemark, der Republik Estland, der Republik Finnland, der Französischen Republik, der Bundesrepublik Deutschland, der Italienischen Republik, der Republik Lettland, der Republik Litauen, des Großherzogtums Luxemburg, der Republik Malta, des Königreichs der Niederlande, der Portugiesischen Republik, der Republik Slowenien und/oder des Königreichs Schweden anzubieten und/oder zu liefern,

wobei das Verfahren zumindest umfasst:

in einer ersten Iteration der Berechnung werden die Vielzahl von Unteraufgaben in einer ersten Verteilung einem der Vielzahl von Rechenknoten (20) und Booster-Knoten (22) zugewiesen und von diesen verarbeitet; dadurch gekennzeichnet, dass

Informationen, die sich auf die Verarbeitung der Vielzahl von Unteraufgaben durch die Vielzahl von Rechenknoten (20) und Booster-Knoten (22) beziehen, verwendet werden, um eine weitere Verteilung der Unteraufgaben zwischen den Rechenknoten (20) und Booster-Knoten (22) zu erzeugen, um sie in einer weiteren Recheniteration zu verarbeiten,

– mittelbare Verletzung von Anspruch 1 des EP 3 743 812 B1 –

insbesondere, wenn bei dem Verfahren

jede der mehreren Unteraufgaben mehr oder weniger für die Verarbeitung durch einen oder mehrere der Rechenknoten oder einen oder mehrere der Booster-Knoten geeignet sein kann;

– mittelbare Verletzung von Anspruch 1 des EP 3 743 812 B1
in der Fassung nach Hilfsantrag 1 –

und/oder, wenn bei dem Verfahren

ein Ressourcenmanager die Zuweisung von Aufgaben und Unteraufgaben zu den Rechenknoten und Booster-Knoten für die erste Iteration in Abhängigkeit von der Rechenaufgabe bestimmt;

– mittelbare Verletzung von Anspruch 1 des EP 3 743 812 B1
in der Fassung nach Hilfsantrag 2 –

und/oder, wenn bei dem Verfahren

die erste Verteilung von Unteraufgaben zwischen Rechenknoten und Booster-Knoten auf einer Skalierbarkeit der Unteraufgabe basiert;

– mittelbare Verletzung von Anspruch 1 des EP 3 743 812 B1
in der Fassung nach Hilfsantrag 3 –

und/oder, wenn bei dem Verfahren

ein Ressourcenmanager auf der Grundlage der Informationen die Zuordnung der Rechenknoten und Booster-Knoten zueinander während der Berechnung der Rechenaufgabe dynamisch ändert;

– mittelbare Verletzung von Anspruch 1 des EP 3 743 812 B1
in der Fassung nach Hilfsantrag 4 –

und/oder, wenn bei dem Verfahren

ein Anwendungsmanager die Informationen empfängt und die weitere Verteilung bestimmt, und wobei ein Ressourcenmanager die Zuweisung von Aufgaben und Unteraufgaben zu den Rechenknoten und Booster-Knoten für die erste Iteration in Abhängigkeit von der Rechenaufgabe bestimmt und wobei der Anwendungsmanager die Informationen empfängt und sie als Eingabe für den Ressourcenmanager verarbeitet, so dass der Ressourcenmanager die weitere Verteilung während der Berechnung der Rechenaufgabe dynamisch ändert.

– mittelbare Verletzung von Anspruch 1 des EP 3 743 812 B1
in der Fassung nach Hilfsantrag 5 –

und/oder, wenn bei dem Verfahren

ein Anwendungsmanager die Informationen empfängt und die weitere Verteilung bestimmt, und wobei ein Ressourcenmanager die Zuweisung von Aufgaben und Unteraufgaben zu den Rechenknoten und Booster-Knoten für die erste Iteration in Abhängigkeit von der Rechenaufgabe bestimmt und wobei der Anwendungsmanager die Informationen empfängt und sie als Eingabe für den Ressourcenmanager verarbeitet, so dass der Ressourcenmanager die weitere Verteilung während der Berechnung der Rechenaufgabe dynamisch ändert, und wobei der Ressourcenmanager die Informationen empfängt, so dass der Ressourcenmanager die Zuordnung der Berechnungsknoten und Booster-Knoten zueinander während der Berechnung der Berechnungsaufgabe dynamisch ändert.

– mittelbare Verletzung von Anspruch 1 des EP 3 743 812 B1
in der Fassung nach Hilfsantrag 6 –

II. Die Beklagten werden weiter verurteilt, innerhalb eines Zeitraums von 30 Tagen nach der Zustellung der Mitteilung i.S.v. R. 118 (8) S. 1 VerfO und gegebenenfalls der beglaubigten Übersetzung,

den Klägerinnen darüber Auskunft zu erteilen, in welchem Umfang sie seit dem 2. Januar 2021 die in Ziffer I. bezeichneten Handlungen begangen hat, und zwar in Form einer für jeden Monat eines Kalenderjahres und nach den in Ziffer I. bezeichneten Erzeugnissen strukturierten Aufstellung der nachfolgenden Informationen:

a) Ursprung und Vertriebswege der verletzenden Erzeugnisse;

b) die erzeugten, hergestellten, ausgelieferten, erhaltenen oder bestellten Mengen und die Preise, die für die verletzenden Erzeugnisse gezahlt wurden;

c) die Identität aller an der Herstellung oder dem Vertrieb der verletzenden Erzeugnisse beteiligten dritten Personen;

d) die Anzahl und die Daten der angebotenen Erzeugnisse;

e) die durchgeführte Werbung, aufgeschlüsselt nach Werbeträgern, ihre Verbreitung, den Vertriebszeitraum und das Vertriebsgebiet; einschließlich der Nachweise für diese Werbetätigkeiten;

f) die Kosten, aufgeschlüsselt nach einzelnen Kostenfaktoren und den erzielten Gewinnen,

wobei zum Nachweis der Angaben die entsprechenden Kaufbelege (nämlich Rechnungen, hilfsweise Lieferscheine) in Kopie vorzulegen sind, wobei geheimhaltungsbedürftige Details außerhalb der auskunfts- und mitteilungspflichtigen Daten geschwärzt werden dürfen;

III. Die Beklagten werden weiter verurteilt,

1. im Falle jeder Zuwiderhandlung gegen die Anordnung gemäß dem Antrag zu Ziff. I. ein wiederholtes Zwangsgeld in Höhe von mindestens 1.000,00 EUR pro verletzendes Erzeugnis;

2. im Falle jeder Zuwiderhandlung gegen die Anordnung gemäß dem Antrag zu Ziff. II. ein wiederholtes Zwangsgeld von mindestens 250,00 EUR pro Tag für jeden Tag der Zuwiderhandlung an das Gericht zu zahlen.

IV. Die Beklagten werden verurteilt, den Klägerinnen EUR 100.000,00 als vorläufigen Schadensersatz zu zahlen, der angepasst wird, wenn die unter Ziffer I. genannten Handlungen fortgesetzt werden.

V. Es wird festgestellt,

1. dass die Beklagten verpflichtet sind, den Klägerinnen für die zu I. bezeichneten und in der Zeit vom 2. Januar 2021 bis zum 1. September 2023 begangenen Handlungen eine angemessene Entschädigung zu zahlen;

2. dass die Beklagten verpflichtet sind, den Klägerinnen allen Schaden zu ersetzen, der ihnen durch die zu I. bezeichneten, seit dem 2. September 2023 begangenen Handlungen entstanden ist und noch entstehen wird.

VI. Die Beklagten werden weiter verurteilt, die Kosten des Rechtsstreits zu tragen.

14 Die Beklagten beantragen,

I. Die Klage wird abgewiesen.

hilfsweise:

Ia. Die Zahlung eines vorläufigen Schadensersatzes nach R. 119 VerfO ist von einer Sicherheitsleistung der Klägerinnen (durch Hinterlegung oder Bankbürgschaft) an die Beklagten in Höhe des zugesprochenen vorläufigen Schadensersatzes abhängig (R. 352 VerfO).

II. Die Klägerinnen tragen die Kosten des Verfahrens als Gesamtschuldner.

15 Mit ihrer Replik vom 13. Juni 2025 haben die Klägerinnen außerdem erstmals eine unmittelbare Verletzung des Streitpatents gestützt auf den Patentanspruch 1 geltend gemacht und entsprechende Anträge angekündigt. Die Zulassung dieser Klageerweiterung wurde mit Anordnung vom 15. Januar 2026 abgelehnt.

16 Die Beklagten haben eine Widerklage auf Nichtigerklärung erhoben. Auf Nachfrage in der mündlichen Verhandlung, ob die Beklagten eine Entscheidung über die Widerklage wünschen, wenn das Gericht bereits anderweitig zu dem Ergebnis kommt, dass das

Streitpatent nicht verletzt sein sollte, haben die Beklagten erklärt, sie seien mit einer solchen Bedingung für die Nichtigkeitswiderklage einverstanden. Sie beantragen insofern:

I. Das Patent EP 3 743 812 wird im Umfang des Anspruchs 1 wie erteilt sowie im Umfang des Anspruchs 1 des jeweiligen Hilfsantrags widerrufen.

II. Die Klägerinnen tragen die Kosten des Widerklageverfahrens als Gesamtschuldner.

17 Die Klägerinnen haben erklärt, dass sie mit Blick auf andere mögliche Verletzer Interesse an einer Entscheidung über die Widerklage haben und beantragen,

1. die Widerklage der Beklagten auf Nichtigerklärung des Europäischen Patents EP 3 743 812 abzuweisen;

   hilfsweise

   das Europäische Patent EP 3 743 812 in der Fassung eines der Hilfsanträge 1 bis 6 gemäß Anlage K 60 (hilfsweise in dieser Reihenfolge) aufrecht zu erhalten; und

2. den Beklagten die Kosten der Widerklage aufzuerlegen.

18 Im Rahmen der Widerklage auf Nichtigerklärung haben die Beklagten nach Ablauf der Schriftsatzfristen mit Schriftsatz vom 24. Januar 2026 die Publikation „Efficient CPU-GPU cooperative computing for solving the subset-sum problem“ von Wan et al. (Anlage BP CR 11) als neuen Stand der Technik vorgelegt. Die beantragte Zulassung des Schriftsatzes hat der Berichterstatter am 30. Januar 2026 mit Stellungnahmefrist für die Klägerinnen zwecks inhaltlicher Auseinandersetzung gewährt. Die Entscheidung über die Zulassung des weiteren Stands der Technik und der damit verbundenen Klageerweiterung hat er dem Spruchkörper vorbehalten.

19 Insofern beantragen die Beklagten,

die Einführung des weiteren Stands der Technik (Dokument Wan et al.) in das Verfahren zuzulassen.

20 Die Klägerinnen beantragen,

die Klageerweiterung nicht zuzulassen.

STREITPUNKTE DER PARTEIEN

Klagebefugnis und Anspruchsberechtigung

21 Die Klägerinnen sind der Ansicht, die zwischen ihnen und mit der FL Systems AG & Co. KG geschlossenen Lizenzverträge seien wirksam. Ein Scheingeschäft hätten die Beklagten nicht dargelegt. Ein verbotenes Insichgeschäft liege nicht vor, weil Herr Frohwitter ausweislich der vorgelegten Handelsregisterauszüge für alle Gesellschaften vom Verbot der Mehrfachvertretung befreit sei. Die Lizenzverträge deckten den vollständigen Schutz- und Anwendungsbereich des Klagepatents ab, so dass die Klägerin zu 1) am gesamten Gegenstand des Streitpatents lizenziert sei. Soweit dies für die Klageberechtigung oder etwaige Ansprüche der Klägerin zu 2) überhaupt notwendig sei, partizipiere die Klägerin zu 2) als Komplementärin der FL Systems AG & Co. KG an der Lizenzvergabe auch wirtschaftlich.

22 Die Beklagten sind der Ansicht, die beiden von der Klägerin zu 2) mit der Klägerin zu 1) und der FL Systems AG & Co. KG geschlossenen Lizenzverträge seien unwirksam. Es handele sich um ein Scheingeschäft, weil eine Gegenleistung nicht vereinbart worden sei. Den gegen alle Lizenzverträge erhobenen Einwand der Mehrfachvertretung haben die Beklagten zuletzt fallengelassen.

23 Die Beklagten sind der Auffassung, angesichts der eingeräumten Lizenzen fehle der Klägerin zu 2) die Prozessführungsbefugnis, jedenfalls aber die Anspruchsberechtigung. Denn für letztere müsse vorgetragen werden, dass sie Inhaberin der von ihr geltend gemachten Ansprüche sei. Bei mehreren Prozessführungsberechtigten, die denselben Anspruch geltend machen, müssten alle Kläger Inhaber dieses Anspruchs für den geltend gemachten Zeitraum sein. Das sei hier aber nicht der Fall. Die Klägerin zu 2) habe sich durch die Lizenzeinräumungen aller Nutzungsmöglichkeiten begeben.

24 Entschädigungsansprüche und darauf basierende Auskunftsansprüche für die Zeit bis 1. September 2023 schieden bereits deshalb aus, weil dazu nicht vorgetragen sei und die Voraussetzungen nicht vorlägen. So existiere bereits nicht die nach französischem und deutschem Recht erforderliche Übersetzung der Patentschrift in die jeweilige Landessprache. Zudem ziehe eine mittelbare Verletzung ohnehin keine Entschädigungsansprüche nach sich. Weitere Ansprüche auf Unterlassung, Schadensersatz und Auskunft fielen aufgrund der ausschließlichen Lizenzen ebenfalls aus, spätestens aber mit der Lizenz an die FL Systems AG & Co. KG. ab dem 21. Dezember 2023. Die Klägerin zu 2) sei von einer Patentverletzung auch nicht betroffen, da sie sich in den Verträgen keine Nutzungsmöglichkeiten vorbehalten habe. Eine wirtschaftliche Partizipation der Klägerin zu 2) an der Lizenzvergabe bestreiten die Beklagten. Dies folge ihrer Ansicht nach auch nicht zwingend aus der Stellung der Klägerin zu 2) als Komplementärin der FL Systems AG & Co. KG. Typischerweise würden nur die Kommanditisten an den Einnahmen einer KG partizipieren.

25 Die Klägerin zu 1) sei frühestens ab dem 7. Juni 2024 und dem Abschluss des Lizenzvertrages mit der FL Systems AG & Co. KG anspruchsberechtigt.

Auslegung

26 Die Klägerinnen sind der Auffassung bei den Rechen- und Booster-Knoten im Sinne des Klagepatents müsse es sich nicht zwingend um autonome oder unabhängige Recheneinheiten handeln. Wenn man zur Auslegung des Patentanspruchs die im Streitpatent erwähnte WO 2012/049247 heranziehen wolle, dann müsse berücksichtigt werden, dass sie sowohl einzelne Prozessoren als Knoten offenbare und zudem die dortige Anordnung von Rechen- und Booster-Knoten darauf ausgelegt sei, Virtualisierung auf allen in Frage kommenden Ebenen zu ermöglichen. Infolgedessen könne ein "Rechenknoten" auch einfach ein Multi-Core-Prozessor sein und ein "Booster-Knoten" eine Grafikprozessor-Einheit (GPU), ein Many-Core-Prozessor oder ein Cluster aus Many-Core-Prozessoren.

27 Rechen- und Boosterknoten seien demzufolge unterschiedlich ausgestaltet und hätten unterschiedliche technische Fähigkeiten. Funktionell seien Rechenknoten und Booster-Knoten zwar unterschiedlich aber sich ergänzend geeignet, heterogene Aufgaben mit verschiedenartigen Aufgaben bzw. Unteraufgaben zu berechnen. Die Verarbeitung der Rechenaufgabe erfolge durch eine geeignete Verteilung der Aufgaben auf Rechen- und Boosterknoten.

28 Die Klägerinnen sind weiterhin der Ansicht, dass der Begriff der Rechenaufgabe mit ihren Unteraufgaben nicht mit einem (homogenen) Datensatz gleichgesetzt werden könne. Eine Rechenaufgabe und ihre Unteraufgaben umfassten Instruktionen oder Anweisungen, die anders als einfache Datensätze nicht ohne weiteres teilbar und infolgedessen regelmäßig heterogen seien.

29 Was die Verteilung der Unteraufgaben angehe, setze Anspruch 1 des Streitpatents nach Ansicht der Klägerinnen eine unmittelbare Verteilung voraus, schon um qualifizierte Informationen gewinnen zu können. Das wäre bei einer mittelbaren Verteilung, in die die Rechenknoten eingeschaltet wären, nicht möglich.

30 Soweit Anspruch 1 des Streitpatents verlange, dass eine weitere Verteilung der Unteraufgaben erfolgt, um sie in einer weiteren Iteration zu verarbeiten, muss es sich nach Auffassung der Klägerinnen nicht exakt um die Unteraufgaben handeln, die Gegenstand der ersten Verteilung gewesen seien. Es gehe vielmehr um die Unteraufgaben, die Gegenstand der weiteren Iteration seien. Es müsse sich nicht zwingend um denselben Satz von Unteraufgaben wie in der ersten Iteration handeln. Die Anzahl der Unteraufgaben könne sich verändern. Unteraufgaben, die Gegenstand der zweiten Verteilung seien, könnten auch neu definiert sein und müssten nicht bereits Gegenstand der ersten Verteilung gewesen sein. Weiterhin stelle die „weitere Recheniteration“ nicht notwendig die zweite Iteration nach der ersten Iteration dar. Schließlich sei in der weiteren Recheniteration auch nicht erforderlich, dass über alle Unteraufgaben hinweg eine andere Verteilung der Unteraufgaben als in der ersten Iteration vorliege. Es sei auch möglich, nur einzelne der Unteraufgaben auf andere Knoten zu verteilen.

31 Der Begriff der „Recheniteration“ ist nach Auffassung der Klägerinnen von einer bloßen Iteration zu unterscheiden und umfasse einen Bearbeitungsschritt bzw. eine Bearbeitungszeit, nach dem oder nach der eine Bewertung des Status des Systems und der ersten Verteilung sinnvollerweise vollzogen werden könne und aussagekräftige Informationen für eine weitere, grundsätzlich systemweite Verarbeitung bzw. Verteilung vorhanden seien. Anspruch 1 des Streitpatents setze nicht voraus, dass die Verarbeitung der Teilaufgaben in der ersten Iteration abgeschlossen werde. Rechenaufgaben für Hochleistungs-Cluster-Computersysteme seien in der Regel für eine Verarbeitung über Tage, Wochen und Monate angelegt. Sie könnten währenddessen auch pausiert oder ausgesetzt werden, müssten aber nicht zwingend abgeschlossen werden. Dies sei aber auch nicht ausgeschlossen, so dass Teilaufgaben in einer weiteren Iteration dann mit anderen Ergebnissen wiederholt werden könnten.

32 Was die Verteilung der Unteraufgaben angehe, könne die erste Verteilung als Funktion der Rechenaufgabe – wie etwa ihre Skalierbarkeit – erfolgen, sei darauf aber nicht beschränkt. Sie könne auch unter anderen Gesichtspunkten wie dem Status und der Verfügbarkeit von Ressourcen, der Abhängigkeit von Unteraufgaben untereinander oder dergleichen erfolgen.

33 Bei den Informationen, auf deren Grundlage nach der Lehre des Streitpatents die weitere Verteilung erfolgen soll, könne es sich nach Auffassung der Klägerinnen um Informationen bezüglich der Verarbeitung der Unteraufgaben, gegebenenfalls um Informationen zur Skalierbarkeit oder um Statusinformationen der Rechen- und Boosterknoten handeln. Dass letzteres bereits im Stand der Technik für das Load-Balancing bereits bekannt gewesen sei, sei unerheblich. Das Streitpatent erwähne beide Arten von Informationen und unterscheide sich vom Stand der Technik („load balancing“) bereits dadurch, dass im Stand der Technik nicht erkannt worden sei, dass das Verarbeiten der Teilaufgaben weitergehenden Einfluss auf den Status der Knoten habe und nicht nur deren Auslastung betreffe, wenn etwa die Eignung des Knotens für eine Rechenaufgabe betrachtet werde. Das Streitpatent unterscheide sich davon durch die Verwendung qualifizierter Informationen, weil die Information über die reine Auslastung der Knoten zu grob granuliert sei.

34 Die Beklagten sind der Auffassung, die Begriffe „Rechenknoten“ und „Booster-Knoten“ in Anspruch 1 des Streitpatents seien wie in dem Stand der Technik WO 2012/049247 zu definieren, auf dem der Gegenstand des Streitpatents ausdrücklich aufbaue. Gemäß der erteilten Fassung EP 2 628 080 dieses Standes der Technik sei ein Knoten in einem Cluster eine eigenständige/unabhängige Recheneinheit, zum Beispiel in Form eines „Stand-alone computer“. Dies ergebe sich im Übrigen auch unmittelbar aus der WO-Schrift. Ferner handele es sich bei den Booster-Knoten um separate Einheiten, die nicht Teil der Rechenknoten seien. Soweit die Klägerinnen meinen, Rechen- und Booster-Knoten seien für verschiedene Unteraufgaben unterschiedlich geeignet, vertreten die Beklagten die Auffassung, dass dies möglich sei, Anspruch 1 des Streitpatents aber nicht darauf beschränkt sei. Auch funktional seien keine Gründe für unterschiedliche Eigenschaften erkennbar.

35 Was die (Unter-)Aufgaben angehe, müsse es sich nicht um „heterogene (Unter-)Aufgaben“ handeln, insbesondere nicht um (Unter-)Aufgaben, die sich dahingehend unterscheiden ließen, ob sie besser auf Rechenknoten oder besser auf Booster-Knoten zu bearbeiten seien. Dies habe keinen Eingang in Anspruch 1 gefunden, vielmehr deute die Beschreibung des Klagepatents auf anderes hin.

36 Hinsichtlich der Zuweisung der Unteraufgaben sind die Beklagten der Auffassung, dass diese nicht zwingend auf der Ebene der Verteilung, mithin direkt erfolgen müsse, insbesondere nicht durch eine zentrale Einheit. Die Entscheidung über die Zuweisung an einzelne Knoten könne auch durch Rechenknoten getroffen werden, die Unteraufgaben an mit ihnen verbundene Booster-Knoten (weiter-)verteilen.

37 Anspruch 1 des Streitpatents verlange nach der ersten Verteilung der Vielzahl von Unteraufgaben eine davon verschiedene zweite Verteilung. Aus dem Wortlaut des Anspruchs ergebe sich jedoch, dass sämtliche Unteraufgaben, die die Rechenaufgabe umfasse, Teil der ersten und der zweiten Iteration sein müssten. Der gleiche Satz an Unteraufgaben müsse in der ersten und zweiten Iteration verarbeitet werden. Die mit der Umverteilung verbundene Effizienzsteigerung trete nur bei der Verwendung der gleichen Unteraufgaben in den mehreren Iterationen ein.

38 Zudem ergebe sich aus dem Begriff der „Iteration“, dass jede der Unteraufgaben mindestens zweimal berechnet werden müsse, nämlich in einer ersten und einer zweiten Iteration. Denn der Begriff bezeichne die wiederholte Berechnung desselben Satzes von Unteraufgaben.

39 Die im Patentanspruch 1 genannten Informationen, die sich auf die Verarbeitung der Vielzahl von Unteraufgaben durch die Vielzahl von Rechen- und Boosterknoten beziehen, seien von Informationen betreffend die Auslastung oder den Status der Rechenknoten und Booster-Knoten abzugrenzen. Anspruchsgemäß seien nur solche Informationen, die – wie etwa die Skalierbarkeit von Unteraufgaben – einen Bezug zur tatsächlichen, ersten Verarbeitung der Unteraufgaben aufweisen, denn erfindungsgemäß soll das System aus diesen Informationen für die zweite/weitere Iteration, also der erneuten Verarbeitung der gleichen Unteraufgaben, lernen. Nicht erfindungsgemäß, sondern vorbekannt sei die Verwendung von Statusinformationen der Knoten für die weitere Verteilung. Die von den Klägerinnen herangezogenen Informationen über die Ressourcenverfügbarkeit bzw. die Auslastung/Kapazität von Knoten stellten im Übrigen nichts anderes als Status-Informationen dar, die nicht anspruchsgemäß seien.

Angegriffene Ausführungsform

40 Die Klägerinnen haben in der Klageschrift erklärt, bei der mit der Klage angegriffenen Ausführungsform handele es sich um DGX-System-Produkte, die mit der Run:ai-Software geliefert werden. Auf das Bestreiten der Beklagten, dass die Software nicht vorinstalliert sei, haben die Klägerinnen in der Replik ihre Behauptung aufrechterhalten, dass die DGX-System-

Produkte mit Run:ai-Software angeboten und geliefert würden. Zugleich haben sie erklärt, selbst wenn Run:ai nicht im "Bundle" mit den angegriffenen DGX-Systemen angeboten und geliefert werde, sondern separat zu erwerben und zu installieren sei, könne an einer mittelbaren Verletzung kein Zweifel bestehen.

41 Insofern behaupten die Klägerinnen, die Nvidia DGX Systeme würden standardmäßig mit der Software des Unternehmens Run:ai ausgeliefert. In der Runa:ai Documentation Library heiße es insofern: „NVIDIA DGS comes bundled out of the box with Run:ai." (Anlage K 21). Die Software sei für den Einsatz auf Clustern der DGX-Systeme vollständig getestet und zertifiziert. Selbst wenn Run:ai vom Endnutzer auf den erworbenen DGX-Systemen installiert werden müsse, stelle Nvidia detaillierte Anweisungen dafür zur Verfügung und integriere die Installation von Run:ai sogar in seine eigene Base Command Manager-Software ("BCM"), die auf allen DGX-Systemen installiert sei (Anlage K 52).

42 Letztlich ist es nach Ansicht der Klägerinnen unerheblich, ob der Kunde die Run:ai-Software selbst installieren müsse, weil in einem komplexen High-Performance-Computing-System, das für Anwendungen wie maschinelles Lernen eingesetzt werde, die Installation einer ganzen Reihe verschiedener Softwareprogramme unerlässlich sei. Eine etwaige Notwendigkeit der gesonderten Installation von Run:ai auf den DGX-Systemen sei ohne Bedeutung für die Frage, ob die Beklagten DGX-Systeme zusammen mit Run:ai als "Bündel" bzw. "Paket" vertreiben. Im Jahr 2023 seien DGX BasePod-Geräte tatsächlich mit der Run:ai-Software angeboten worden. Nvidia stelle auch den First-Level-Support für Run:ai bereit und habe Run:ai in den Software-Stapel des DGX-Cloud Create-Dienstes integriert. Nach alledem sei die Run:ai-Software – unabhängig von der Vertriebs- oder Installationsmethode – ein wesentlicher Bestandteil des Bündels von Nvidia und daher ein integraler Bestandteil ihres Angebots.

43 Die Beklagten behaupten, beim Kauf eines DGX-Produkts sei die Run:ai-Software noch nicht installiert. Die Run:ai-Software müsse von den Kunden der Beklagten oder jemandem im Auftrag der Kunden eigenständig heruntergeladen und installiert werden. Das gelte auch für die Zeit nach der Übernahme von Run:ai durch Nvidia. Etwas anderes ergebe sich auch nicht aus der Software-Dokumentation (Anlage K 21). Darin werde auch ausgeführt, dass Run:ai erst nach dem Erwerb der DGX-Hardware darauf zu installieren und einzurichten sei. Die Beschreibung „bundled out of the box with Run:ai" meine schlicht, dass die DGX-Produkte so gestaltet seien, dass die Run:ai-Software von den Kunden der Beklagten darauf installiert und verwendet werden könnten, aber nicht darauf vorinstalliert sei. Denn die Run:ai-Software sei für ein ordnungsgemäßes Funktionieren der DGX-Systeme nicht erforderlich. Um die Run:ai-Software installieren zu können, müsse der Käufer des DGX-Systems ein separates Software-Abonnement erwerben

44 Die Beklagten sind der Ansicht, dass sich die Klage ausschließlich gegen DGX-Hardwareprodukte der Beklagten richte, die vermeintlich zusammen mit der Run:ai-Software

geliefert würden. Solche Ausführungsformen gebe es nicht. Soweit die Klägerinnen nun versuchten, den Gegenstand der angegriffenen Ausführungsform zu erweitern, sei eine solche Erweiterung unzulässig, da sie sich an den Voraussetzungen von Regel 263 VerfO messen lassen müsse. Dafür fehle ein entsprechender Antrag.

Verletzung

45 Die Klägerinnen sind der Auffassung, die angegriffene Ausführungsform sei – unabhängig davon, ob die Run:ai-Software vorinstalliert sei oder nicht – zur Anwendung des geschützten Verfahrens geeignet. Dies könne exemplarisch anhand des DGX H100 System gezeigt werden.

46 Die angegriffene Ausführungsform stelle ein heterogenes Rechensystem im Sinne des Streitpatents dar. Es umfasse im Fall des DGX H100 Systems eine Vielzahl von Rechenknoten in Form der genutzten Intel Xeon 8480C CPUs und eine Vielzahl von Booster-Knoten in Form der NVIDIA H100 Tensor Core GPUs. Sowohl einzelne DGX-Server (Single Appliance Read) als auch mehrere dieser Server, die mittels eines Hochleistungsnetzwerks zu einem DGX SuperPOD-System zusammengeschaltet seien (Multi Appliance Read), stellten jeweils ein heterogenes Rechensystem im Sinne des Streitpatents dar, ihre Prozessoren die jeweiligen „Nodes“.

47 Nach Auffassung der Beklagten könnten gemäß der Funktionalität der Run:ai-Software, Rechenaufgaben, dort als „workloads“ bezeichnet, aus mehreren Unteraufgaben, auch „pods“ genannt, bestehen, wobei jeder Pod auf einem eigenen Node – CPU oder GPU – ausgeführt werde. Der Run:ai-Scheduler treffe hierbei Entscheidungen für die Ressourcenzuweisung anhand vorab festgelegter Regeln. Werde Workload zur Verarbeitung eingereicht, überprüfe der Run:ai-Scheduler, ob für diese Aufgabe genügend Ressourcen zur Verfügung stehen, indem er die Anfrage mit den aktuell zugeteilten Ressourcen und den festgelegten Quotas vergleiche. Mithin würden in einer ersten Iteration der Berechnung die Vielzahl von Unteraufgaben in einer ersten Verteilung einem der Vielzahl von Rechen- und Boosterknoten zugewiesen und von diesen verarbeitet.

48 Der Scheduler berechne jedoch die faire Ressourcenzuteilung für jedes Projekt ständig neu und verteile die Ressourcen entsprechend neu (Anlage K 40). Die Run:ai-Software könne die Workloads im laufenden Betrieb dynamisch verkleinern und erweitern, was Elastizität genannt werde. Wenn zusätzliche Ressourcen verfügbar seien, ermögliche es die Elastizitätsfunktion, diese während der Laufzeit den Workloads hinzuzufügen, um die Ausführung von Workload zu beschleunigen. Mithin würden Informationen, die sich auf die Verarbeitung der Unteraufgaben beziehen, verwendet, um eine weitere Verteilung der Unteraufgaben zwischen den Knoten zu erzeugen, um sie in einer weiteren Recheniteration zu verarbeiten.

49 Der Run:ai-Scheduler ordne die Ressourcen dem Workload nicht nur nach ihrer Verfügbarkeit, sondern auch nach den Quoten- und Fairness-Prinzipien von Run:ai zu. Dies habe zur Folge,

dass der Run:ai Scheduler in bestimmten Szenarien Ressourcen gemäß diesen Fairness-Prinzipien von einem Workload auf einen anderen umverteile. Habe der Scheduler einem Workload in einem Planungszyklus mehr als die erforderlichen Ressourcen zugewiesen, könne er die überzähligen Ressourcen in einem weiteren Planungszyklus einem anderen Workload zuweisen, der Ressourcen erfordert. Nach Auffassung der Klägerinnen stellen die Ressourcen, die einem Workload in einem vorherigen Planungszyklus tatsächlich zugewiesen wurden, Informationen dar, die sich auf die Verarbeitung der Teilaufgaben beziehen. Denn auch der Status der Knoten und allgemein, wie eine Verteilung aufgrund der Situation im DGX-Gerät besser gestaltet werden könne, stellten Informationen in Bezug auf die Berechnung der Unteraufgaben im Sinne des Streitpatents dar, weil sie auf die weitere Verteilung der Aufgaben auf die Knoten Einfluss nehmen könnten.

50 Die Klägerinnen sind der Ansicht, die Berücksichtigung patentgemäßer Informationen für die weitere Verteilung von Unteraufgaben für eine weitere Recheniteration ergebe sich auch aus einer Funktion der Run:ai-Software, die „Hugging Face“-Inferenzmodelle unterstütze. Demnach erstelle Run:ai jedes Mal, wenn eine variable Bedingung in Bezug auf den zugrundeliegenden Inferenz-Workload (d.h. die Menge der Teilaufgaben) erfüllt sei, einen neuen Replica-Pod (d.h. eine neue Unteraufgabe) und weise diesem Ressourcen zu. Die entsprechenden Variablen Latenz, Durchsatz und Parallelität stellten Informationen in Bezug auf die Verarbeitung der Unteraufgaben dar, aus denen sich der Inferenz-Workload zusammensetze.

51 Nach Auffassung der Klägerinnen stelle jeder Planungszyklus des Run:ai-Schedulers eine Iteration im Sinne des Streitpatents dar. Davon gebe es mehrere. Die Berechnung des Workload könne unterbrochen und währenddessen die dafür zugewiesenen Ressourcen geändert werden. Diese „Unterbrechungen“ bezeichneten den Planungszyklus des Run:ai-Schedulers. Dies gelte für das Single Appliance Read ebenso wie für das Multi-Appliance Read. Es sei nicht notwendig, dass die Teilaufgaben innerhalb jeder Iteration abgeschlossen werden.

52 Die Beklagten sind der Auffassung, die angegriffene Ausführungsform stelle kein anspruchsgemäßes heterogenes Rechensystem mit einer Vielzahl an Rechen- und Booster-Knoten dar. Es handele sich vielmehr um einen einzelnen Server, das heißt einen einzelnen „node“, der CPUs und GPUs enthalte, nicht aber eigenständige Rechen- oder Booster-Knoten wie vom Anspruch 1 gefordert. CPUs und GPUs seien keine eigenständigen Recheneinheiten im Sinne des Streitpatents, insbesondere die GPUs seien keine Booster-Knoten, sondern Komponenten eines DGX-Systems und damit eines Rechenknotens. Gemäß der Run:ai-Dokumentation handele es sich bei Knoten um eigenständige Server, die verschiedene Komponenten wie CPU und GPU umfassten, die aber selbst keine Knoten seien.

53 Die Beklagten teilen die Auffassung, die von der Run:ai-Software genannten „workloads“ könnten als Rechenaufgaben und die „pods“ als Unteraufgaben angesehen werden. Die

Beklagten behaupten jedoch, dass [redacted]

54 Soweit der Scheduler der Run:ai-Software für eine dynamische Ressourcenzuweisung sorge, wobei das „elasticity feature“ die Verwendung von zusätzlichen Ressourcen – sofern verfügbar – für die Aufgabenverarbeitung ermögliche, werde die Lehre des Streitpatents gerade nicht verwirklicht. Die Zuweisung von Workloads zu den CPUs und GPUs der angegriffenen Ausführungsform basiere auf der Verfügbarkeit der Ressource, d.h. dem Status bzw. der Auslastung der Ressourcen. Diese Verfügbarkeit stelle keine Information im Sinne des Streitpatents dar, die sich auf eine Verarbeitung der Unteraufgaben beziehe.

55 Auch die von den Klägerinnen erwähnten „Prioritäten“ seien nach Ansicht der Beklagten keine anspruchsgemäßen Informationen, die sich auf die Verarbeitung der Pods des Workloads beziehen. Sie würden von den Nutzern festgelegt und zudem nur die Priorisierung der Workloads untereinander festlegen, spiele für die Verarbeitung eines einzelnen Workloads aber keine Rolle und noch weniger für die Verarbeitung der Pods.

56 Ebenso wenig betreffe der so genannte „Fairshare“ anspruchsgemäße Informationen, die sich auf die Verarbeitung der Pods eines Workloads beziehen. Der Fairshare gebe lediglich die Summe der garantierten Ressourcen plus den Anteil der nicht garantierten Ressourcen eines Projekts an. Der Fairshare betreffe daher ebenfalls lediglich Informationen über die Verfügbarkeit von Ressourcen. Die Informationen über Ressourcen, die einem Workload in einem vorherigen Planungszyklus tatsächlich zugewiesen wurden, stellten keine Informationen dar, die sich auf die Verarbeitung der Unteraufgaben beziehen, sondern um Informationen betreffend die Verfügbarkeit und Auslastung der GPU. Gleiches gelte für Informationen über Ressourcen, denen derzeit keine Pods zugewiesen seien.

57 Was den „Hugging-Face inference workload” angehe, sind die Beklagten der Ansicht, dass auch dieser Vortrag schon in der Verletzungsklage hätte erfolgen können. Ungeachtet dessen sei ein „Hugging-Face inference workload“ keine anspruchsgemäße Rechenaufgabe. Vielmehr würden die Pods eines „Hugging-Face inference workloads“ – wie auch normale Workloads – nur einmal berechnet. Es gebe keine weitere Iteration. Außerdem gebe es keine zweite Verteilung. Die „Hugging-Face inference workloads“ basierten auf dem Prinzip, dass bei Bedarf ein oder mehrere neue „replica“ erzeugt würden. Diese Replica seien aber keine anspruchsgemäßen Unteraufgaben, sondern nur eine Kopie des ersten (Teils des) Workloads. Gehe man von einer ersten und zweiten Iteration von Replica aus, würden nicht die gleichen Pods umverteilt. Vielmehr handele es sich dann um Iterationen unterschiedlicher und nicht mehr gleicher Pods, weil die Replika neu hinzugekommen seien und nun andere Pods bearbeitet würden. Das bloße Hinzufügen von Replica stelle ebenfalls keine Umverteilung dar, weil sich die Verteilung der übrigen Pods nicht ändere.

58 [redacted]. Eine Unterbrechung der Berechnung der Pods finde nicht nach einer vollständigen Berechnung/Iteration der Pods statt. Zudem erfolge die Berechnung der Pods lediglich anhand eines festen Planungszyklus. Dieser sei unabhängig von der Berechnung der jeweiligen Workloads oder Pods, sondern laufe global ab.

59 Mangels weiterer Iterationen, so die Beklagten, finde auch keine Änderung der Zuweisung, d.h. Umverteilung der Pods, zwischen zwei Iterationen statt.

60 Schließlich vertreten die Beklagten die Ansicht, dass eine zweite Verteilung der Unteraufgaben (Pods) einer Rechenaufgabe (Workload) schon deshalb nicht stattfinde, weil die Unteraufgaben der ersten Verteilung nicht die gleichen Unteraufgaben seien, die Gegenstand der angeblich zweiten Verteilung seien. Pods eines Workloads, denen in einem ersten Planungszyklus Ressourcen zugewiesen worden seien, die aber in einem weiteren Planungszyklus anderen Workloads zugewiesen worden seien, würden nicht neu verteilt, sondern ihre Bearbeitung werde unterbrochen und zurückgestellt. Es würden also vor und nach dem zweiten Planungszyklus nicht die „gleichen“ Unteraufgaben bearbeitet.

Verletzungshandlungen und Rechtsfolgen

61 Die Klägerinnen sind der Ansicht, den Beklagten sei unabhängig davon, ob die Run:ai-Software vorinstalliert sei, eine mittelbare Patentverletzung vorzuwerfen. Die DGX-Systeme verwirklichten schon für sich genommen und ohne Run:ai zahlreiche Anspruchsmerkmale und seien somit "Mittel, die sich auf ein wesentliches Element der Erfindung beziehen." Die Vorbereitung für und Abstimmung auf Run:ai belege zudem, dass die DGX-Systeme für eine klagepatentgemäße Verwendung insbesondere mit Run:ai geeignet und bestimmt seien. Die Beklagten wüssten auch, dass die DGX-Systeme zur Anwendung des patentgeschützten Verfahrens bestimmt und geeignet seien und dass der Abnehmer der Systeme diese auch im Geltungsbereich des Streitpatents auf diese Weise verwenden werde.

62 Die Beklagten meinen, es fehle schon an einer mittelbaren Patentverletzung, weil es die angegriffene Ausführungsform – DGX-System-Produkte mit vorinstallierter Run:ai-Software – gar nicht gebe. Es fehle an einem angegriffenen Mittel im Sinne von Art. 36 EPGÜ. Das Handeln der Beklagten zu 1) begründe keine mittelbare Verletzung von Anspruch 1 des Streitpatents. Sie die Beklagte zu 1) stelle nur die DGX-Systeme her. Die Installation der Software erfolge jedoch durch die Abnehmer, die dann auch das geschützte Verfahren anwendeten. Für die Beklagte zu 2) gelte das Gleiche wie für die Beklagte zu 1). Alle angeblichen Verletzungshandlungen der Beklagten zu 2) bezögen sich ausschließlich auf die von der Beklagten zu 1) hergestellten DGX-Produkte. Zudem würden die von den Klägerinnen behaupteten „Verkaufsförderungsaktionen“ und die „Gewährleistung von

Produktzuverlässigkeit und Qualitätsmanagement“ der Beklagten zu 2) im Hinblick auf die DGX-Systeme nicht weiter beschrieben, geschweige denn durch Beweisangebote belegt.

63 Hinsichtlich etwaiger Rechtsfolgen meinen die Beklagten, dass keine starre Frist für die Auskunft festgesetzt werden solle. Auch seien nicht alle beantragten Angaben im Rahmen der Auskunft zu erteilen. Eine Verurteilung zur Vernichtung, Rückruf und/oder Entfernung aus den Vertriebswegen sei unverhältnismäßig. Die Festsetzung von Zwangsgeldern solle nicht von starren Beträgen und Fristen vorab abhängig gemacht werden. Auch seien die Höhe und die Bedingungen für die Anpassung des vorläufigen Schadensersatzes unklar. Jedenfalls hätten die Klägerinnen insoweit vorab Sicherheit zu leisten. Der Antrag auf Feststellung der Schadensersatzpflicht sei unbestimmt, eine Gesamtgläubigerschaft bestehe jedenfalls nicht.

Widerklage auf Nichtigerklärung

64 Die Beklagten halten Anspruch 1 des Streitpatents für nicht patentfähig.

65 Die technische Lehre des Patentanspruchs werde durch die Patentanmeldung US 2012/0233486 A1 („Phull“) neuheitsschädlich vorweggenommen. Insbesondere offenbare Phull die Zuweisung von Unteraufgaben an Booster-Knoten. Insofern sei es unerheblich, wenn der Boosterknoten nicht unmittelbar an der Verteilung der Unteraufgaben teilnehme, sondern einer CPU zugewiesen sei, die ihm Aufgaben abgebe (so genanntes „Offloading“). Ungeachtet dessen offenbare Phull auch eigenständige GPU. Soweit Phull beschreibe, dass Datensätze neu aufgeteilt und zugewiesen würden, handele es sich auch bei der Verarbeitung von Daten um eine Rechenaufgabe im Sinne des Streitpatents.

66 Nach Ansicht der Beklagten ist die technische Lehre des Streitpatents auch nicht neu gegenüber US 2017/0109207 A1 („Li“). Insofern genüge es, wenn Rechenknoten und Booster-Knoten unterschiedliche Rechenleistung hätten, auch wenn sie sonst strukturell gleicher Natur seien. Es würden auch Informationen bezüglich der Verarbeitung der Unteraufgaben im Sinne des Streitpatents verwendet. Eine Abfrage, inwieweit ein Knoten ein Datenvolumen abgearbeitet hat, liefere solche Informationen, die sich mithin auf die Verarbeitung der Vielzahl von Unteraufgaben durch die Vielzahl von Rechenknoten und Booster-Knoten bezögen.

67 Jedenfalls aber sei die Lehre des Streitpatents durch „Li“ in Kombination mit dem Fachwissen oder „Phull“ nahegelegt.

68 Die Beklagten vertreten weiter die Auffassung, der mit der Duplik vom 13. August 2025 vorgelegte weitere Stand der Technik sei in das Verfahren zuzulassen. Zur Begründung führen sie an, der weitere Stand der Technik sei frühestmöglich vorgelegt worden, weil sich die Klägerinnen in ihrer Replik eine Anspruchsauslegung zu Nutze gemacht hätten, die von ihrer früheren Auslegung und der im Erteilungsverfahren vorgenommenen Auslegung abweiche und deswegen in dieser Form nicht zu erwarten gewesen sei. Nach der neuen

Auslegung möchten die Klägerinnen unter Informationen in Bezug auf die Berechnung der Unteraufgaben nun auch Informationen über den Status (d.h. die Verfügbarkeit) von Knoten eines heterogenen Rechensystems verstehen. Auf der Grundlage dieser Auslegung sei die Lehre des Streitpatents angesichts des Artikels „„Method for scheduling data flow task and apparatus“ von Deng nicht neu, jedenfalls aber nahegelegt durch eine Kombination der WO 2912/049247 A1 oder Deng mit der Publikation “Customized Dynamic Load Balancing for a Network of Workstations” von Zaki.

69 Weiterhin meinen die Beklagten, dass die Entgegenhaltung Wan in das Verfahren zuzulassen sei. Diese sei trotz zweier Suchaufträge an zwei verschiedene namhafte Rechercheunternehmen nicht aufgefunden worden. Erst durch die Recherche für einen anderen Mandanten der Kanzlei der Beklagtenvertreter seien sie am 14. Januar 2026 auf diese Entgegenhaltung aufmerksam geworden. Es liege in der Natur der Sache, dass die Ergebnisse einer Recherche nach relevantem Stand der Technik keinen Anspruch auf Vollständigkeit hätten. Das gelte erst recht für wissenschaftliche Publikationen aus einem Nicht-top-Tier-Journal.

70 Wan nimmt nach Auffassung der Beklagten die Lehre des Streitpatents neuheitsschädlich vorweg. Insbesondere müssten nach eigener Auslegung der Klägerinnen die in der zweiten Iteration verarbeiteten Unteraufgaben nicht mit denen der ersten Iteration identisch sein. Unter dieser Annahme sei die Lehre des Streitpatents in Wan offenbart.

71 Die Klägerinnen halten Anspruch 1 des Streitpatents für patentfähig. Die patentgemäße Erfindung sei neu gegenüber „Phull“. Die Entgegenhaltung offenbare weder eine heterogene, in Unteraufgaben unterteilte Berechnungsaufgabe, noch eine im Sinne der beanspruchten technischen Lehre erfolgende Verteilung von Unteraufgaben auf patentgemäß zusammenspielende Rechen- und Booster-Knoten.

72 Li ist nach Ansicht der Klägerinnen ebenfalls nicht neuheitsschädlich. Die Entgegenhaltung offenbare nicht schon dadurch Booster-Knoten, weil sie Rechenknoten unterschiedlicher Leistung beschreibe. Es würde vielmehr ausschließlich die Verwendung von CPU offenbart, die keine Booster seien. Weiterhin finde auch keine anspruchsgemäße Umverteilung von Unteraufgaben statt, sondern lediglich ein Lastenausgleich homogener Datenmengen. Irgendwelche Berechnungen oder Rechenaufgaben seien nicht offenbart, ebenso wenig Informationen bezogen auf die Verarbeitung solcher Aufgaben.

73 Die Lehre des Streitpatents beruhe auch auf einer erfinderischen Tätigkeit. Der Fachmann habe keinen Anlass gehabt, Booster in das von Li offenbart System zu integrieren und Unteraufgaben zwischen den Knoten zu verteilen. Hinsichtlich Phull seien schon in Kombination mit Li nicht alle Merkmale des Patentanspruchs offenbart.

74 Der erstmals mit der Duplik vorgelegte weitere Stand der Technik sei nicht zuzulassen. Dieser sei verspätet und die angeführte Begründung nicht überzeugend. Die Klägerinnen verfolgten von Anfang an eine Auslegung, nach der die angegriffene Ausführungsform patentverletzend sei.

75 Nach Ansicht der Klägerinnen sei auch die Entgegenhaltung Wan nicht in das Verfahren zuzulassen. EPGÜ und Verfahrensordnung nähmen mit dem "front-loaded" System des Verfahrens bewusst in Kauf, dass Stand der Technik, der erst nach Ablauf aller Fristen aufgefunden wird, nicht mehr berücksichtigungsfähig ist. Die Beklagten trügen nichts vor, was eine Abweichung von diesem Regelfall rechtfertigen könnte. Im Übrigen würden die Beklagten durch eine Zulassung unangemessen benachteiligt. In der kurzen Zeit bis zur mündlichen Verhandlung sei eine sachgerechte Verteidigung, die eventuell auch Hilfsanträge umfassen könne, nicht möglich.

76 Wan nehme die Lehre des Streitpatents auch nicht neuheitsschädlich vorweg. Eine „subtask“ nach Wan werde in einer einzigen Iteration vollständig abgearbeitet. Es fehle insofern an einer iterativen Bearbeitung einer Aufgabe. Abgesehen davon finde auch keine erste und weitere Verteilung von Unteraufgaben statt. Es werde lediglich die zu bearbeitende Datenmenge nach jeder Iteration in Abhängigkeit von einem jeweils neu berechneten Aufgabenverteilungsverhältnis angepasst. Ebenso sei eine dynamische Zuweisung von Rechen-Knoten und Boosterknoten nicht offenbart.

GRÜNDE FÜR DIE ENTSCHEIDUNG

77 Die Verletzungsklage ist zulässig, aber unbegründet. Die Widerklage bedarf aufgrund dessen keiner Entscheidung.

A Zulässigkeit der Verletzungsklage

78 Die Verletzungsklage ist zulässig. Insbesondere sind die Klägerinnen gemäß Art. 47 EPGÜ berechtigt, das Einheitliche Patentgericht anzurufen.

I. Klägerin zu 2)

79 Die Klägerin zu 2) ist gemäß Art. 47 Abs. 1 EPGÜ berechtigt, das Einheitliche Patengericht anzurufen, da es sich bei ihr um die Inhaberin des Streitpatents handelt.

80 Dagegen kann nicht mit Erfolg eingewandt werden, die Klägerin zu 2) habe ausschließliche Lizenzen am Streitpatent vergeben mit der Folge, dass sie selbst nicht mehr zur Nutzung des Streitpatents berechtigt und daher nicht in ihren eigenen Rechten betroffen sei (vgl.

Lokalkammer Düsseldorf, Entscheidung vom 13.05.2025, UPC_CFI_505/2024 – Sanofi gg. Amgen).

81 Art. 47 Abs. 1 EPGÜ knüpft die Parteistellung des Klägers allein an die Eigenschaft der Patentinhaberschaft. Dem liegt die Vorstellung zugrunde, dass der Patentinhaber berechtigt sein soll, alle Rechte aus seinem Patent im Klagewege vor dem Einheitlichen Patentgericht geltend zu machen. Daher ist der Patentinhaber auch gemäß Art. 47 Abs. 4 EPGÜ berechtigt, dem von einem ausschließlichen Lizenznehmer angestrengten Verfahren als Partei beizutreten. Es ist keine Voraussetzung, dass dem Patentinhaber in eigener Person Ansprüche durch eine Patentverletzung entstanden sind, sofern nur das eigene Patent betroffen ist. Gleiches lässt sich aus Art. 47 Abs. 2 EPGÜ folgern, wonach der ausschließliche Lizenznehmer nur ein Klagerecht neben dem Patentinhaber hat, wenn dieses nicht sogar vertraglich ausgeschlossen ist. Nach alledem ist der Patentinhaber immer klagebefugt.

II. Klägerin zu 1)

82 Aber auch die Klägerin zu 1) ist klagebefugt. Ihre Berechtigung ergibt sich aus Art. 47 Abs. 2 EPGÜ.

1. Ausschließliche Lizenz

83 Die Klägerin zu 1) ist Inhaberin einer umfassenden ausschließlichen Lizenz am Streitpatent.

a)

84 Die Klägerinnen schlossen am 29. August 2021 einen Vertrag, mit dem die Klägerin zu 2) der Klägerin zu 1) eine ausschließliche Lizenz am Streitpatent einräumte (Anlage K 50).

85 Gemäß Ziffer 11.2 dieses Lizenzvertrages ist auf den Vertrag deutsches Sachrecht anwendbar.

86 Der Vertrag ist wirksam zustande gekommen. Er wurde für beide Seiten jeweils von ihrem Vorstandsvorsitzendem Bernhard Frohwitter unterzeichnet. Dieser war und ist ausweislich der vorgelegten Handelsregisterauszüge für beide Klägerinnen einzelvertretungsberechtigt (Anlagen K 1, K 2 und K 66). Zudem durfte er entgegen § 181 BGB für beide Vertragsparteien zugleich den Vertrag unterzeichnen. Die Beklagten haben zu Recht den Einwand des Insichgeschäfts fallengelassen. Denn aus den vorgelegten Handelsregisterauszügen ergibt sich, dass der Vorstandsvorsitzende Frohwitter für beide Klägerinnen vom Verbot der Selbstkontrahierung befreit ist.

87 Der Lizenzvertrag ist nicht gemäß § 117 Abs. 1 BGB unwirksam. Auch wenn die Einräumung der Lizenz unentgeltlich erfolgte, erfolgte sie nicht zum Schein. Es gibt weder einen Rechtssatz noch nur die Vermutung, dass eine Willenserklärung, mit der der einen Seite eine

Leistung ohne synallagmatische Gegenleistung versprochen wird, nur zum Schein abgegeben wurde. Auch die Beklagten tragen für eine solche Wertung keine weiteren Anhaltspunkte vor. Dass die Erklärungen nicht zum Schein abgegeben wurden, sondern tatsächlich die Rechtsfolge einer wirksamen Lizenzeinräumung gewollt war, ergibt sich auch aus dem Umstand, dass die Klägerin zu 1) im vorliegenden Fall als Klägerin auftritt. Dies kann sie nur, wenn ihr eine ausschließliche Lizenz wirksam eingeräumt wurde.

Der Inhalt des Lizenzvertrages vom 29. August 2021 ist zwischen den Parteien unstreitig. Demnach sollte die ausschließliche Lizenz gemäß Ziffer 4.1 dieses Vertrags sachlich den gesamten Anwendungsbereich der von den Vertragsschutzrechten – darunter die Anmeldung zum Streitpatent – umfassten Erfindungen, insbesondere die Mikroelektronik, umfassen. Von der Lizenz sollte nur der Anwendungsbereich der Systemarchitektur von Supercomputern (High Performance Computing) einschließlich Cloud-Computing und die Einbindung von Quantencomputern in HPC-Umgebungen ausgeschlossen sein.

b)

88 Die Klägerin zu 2) räumte der Klägerin zu 1) weiterhin über die FL Systems AG & Co. KG eine ausschließliche Lizenz am Klagepatent für den übrigen sachlichen Anwendungsbereich ein, so dass die Klägerin zu 1) umfassend ausschließlich lizenziert ist.

aa)

89 Am 21. Dezember 2023 schloss die Klägerin zu 2) mit der FL Systems AG & Co. KG einen Lizenzvertrag, mit der letzterer eine ausschließliche Lizenz am Klagepatent für den Anwendungsbereich der Systemarchitektur von Supercomputern (High Performance Computing) einschließlich Cloud-Computing und die Einbindung von Quantencomputern in HPC-Umgebungen eingeräumt wurde (Anlage K 48).

90 Der Vertrag ist wirksam zustande gekommen. Dies bemisst sich gemäß Ziffer 12.2 des Vertrages nach deutschem Sachrecht. Der Vorstandsvorsitzende der Klägerin zu 2), Bernhard Frohwitter, unterzeichnete den Vertrag für die Klägerin zu 2) als Vertragspartnerin und zugleich für diese als Komplementärin der FL Systems AG & Co. KG. Ausweislich der Handelsregisterauszüge beider Vertragsparteien war Herr Bernhard Frohwitter zu dem Zeitpunkt einzelvertretungsberechtigt für die Klägerin zu 2) bzw. die FL Systems AG & Co. KG und auch vom Verbot der Selbstkontrahierung aus § 181 BGB befreit (Anlagen K 1, 63 und 64). Auch die Beklagten wenden sich nicht dagegen.

91 Soweit die Beklagten einwenden, dieser Lizenzvertrag sei gemäß § 117 Abs. 1 BGB nichtig, da es sich mangels Gegenleistung um ein Scheingeschäft handele, kann dem nicht gefolgt werden. Zur Begründung wird ohne Einschränkung auf die Ausführungen zu dem Lizenzvertrag zwischen den Klägerinnen vom 29. August 2021 Bezug genommen.

bb)
Die FL Systems AG & Co. KG räumte ihrerseits der Klägerin zu 1) mit Vertrag vom 7. Juni 2024 eine ausschließliche Lizenz im Umfang des Vertrages vom 21. Dezember 2023 ein (Anlage K 49). Für diesen Vertrag gelten die Ausführungen wie für die beiden vorangehenden Verträge. Auf ihn ist deutsches Sachrecht anwendbar. Der Vertrag kam durch die Unterzeichnung durch den Vorstandsvorsitzenden der Klägerin zu 1) und der Klägerin zu 2) als Komplementärin der FL Systems AG & Co. KG wirksam zustande, da Herr Bernhard Frohwitter sowohl zur Einzelvertretung berechtigt und vom Verbot der Selbstkontrahierung befreit war (vgl. Handelsregisterauszüge K 2, 63 und 64). Den Einwand der Nichtigkeit aufgrund eines Scheingeschäfts haben die Beklagten zu Recht nicht erhoben. Die Vertragsparteien vereinbarten sogar eine Gegenleistung.

2. Unterrichtung der Patentinhaberin

Zwischen den Parteien ist unstreitig, dass die Klägerin zu 2) als Patentinhaberin von der Verletzungsklage der Klägerin zu 1) im Sinne von Art. 47 Abs. 2 EPGÜ unterrichtet wurde, soweit dies überhaupt erforderlich ist, wenn beide Parteien von vornherein zusammen Klage erheben.

B Begründetheit der Verletzungsklage

92 Die Verletzungsklage ist unbegründet.

I. Stand der Technik, Problem und Lösung

93 Die Erfindung nach dem Streitpatent betrifft einen Mechanismus zur Ausführung von Rechenaufgaben innerhalb einer Rechenumgebung, insbesondere einer heterogenen Rechenumgebung, die für die parallele Verarbeitung einer Rechenaufgabe ausgelegt ist.

94 Das Streitpatent führt zum Stand der Technik aus, dass es sich bei der Erfindung nach dem Streitpatent um eine Weiterentwicklung des in der früheren Anmeldung WO 2012/049247 A1 (nachfolgend: WO'247, in diesem Verfahren vorgelegt als Anlage BP 4) beschriebenen Systems handele, das eine Cluster-Computerarchitektur mit einer Vielzahl von Rechenknoten und einer Vielzahl von Boostern beschreibe, die über eine Kommunikationsschnittstelle miteinander verbunden seien. Ein Ressourcenmanager sei dafür verantwortlich, einen oder mehrere der Booster und die Rechenknoten während der Laufzeit dynamisch einander zuzuweisen. Ein Beispiel für ein solches dynamisches Prozessmanagement sei in Clauss et al. „Dynamic Process Management with Allocation-internal Co-Scheduling towards interaktives Supercomputing", COSH 19. Januar 2016, Prag, CZ, offenbart. Laut Streitpatent biete die in der WO '247 offenbarte Anordnung zwar eine flexible Anordnung für die Zuweisung von Boostern zu Rechenknoten, befasse sich jedoch nicht mit der Frage, wie Aufgaben

zwischen Rechenknoten und Boostern verteilt werden sollen (Abs. [0002]; Absatzangaben ohne Quellenangabe sind solche des Streitpatents, Anlage K 44).

95 Die Anordnung nach WO '247 wird auch von Eicker et al. in „The DEEP Project An alternative approach to heterogeneous cluster-computing in the many core era", Concurrency Computat.: Pract. Exper. 2016; 28:2394-2411, beschrieben. Demnach umfasse das heterogene System eine Vielzahl von Rechenknoten und eine Vielzahl von Booster-Knoten, die durch ein schaltbares Netzwerk verbunden seien. Um eine Anwendung zu verarbeiten, werde die Anwendung „taskifiziert" [„taskified"], um einen Hinweis darauf zu geben, welche Aufgaben von einem Rechenknoten auf einen Booster ausgelagert werden können. Diese Aufgabenteilung wird erreicht, indem der Anwendungsentwickler den Code mit „Pragmas" versieht, die Abhängigkeiten zwischen verschiedenen Aufgaben anzeigen, sowie mit „Labels", die die hochskalierbaren Code-Teile kennzeichnen, die von einem Booster verarbeitet werden sollen. Skalierbarkeit bedeutet in diesem Zusammenhang, dass bei steigender Auslastung des Dienstes durch eine inkrementelle und lineare Erhöhung der Hardware das gleiche Serviceniveau pro Benutzer aufrechterhalten werden kann (Abs. [0003]).

96 Das Streitpatent nennt weiterhin die US 2017/0262319 A1, die in einem Aspekt Laufzeitprozesse beschreibe, die die Verteilung von Daten und die Zuordnung von Aufgaben zu Rechenressourcen vollständig oder teilweise automatisieren könnten. Ein sogenannter Tuning-Experte, d. h. eine menschliche Bedienperson, könne weiterhin erforderlich sein, um Aktionen den verfügbaren Ressourcen zuzuordnen. Im Wesentlichen werde beschrieben, wie eine bestimmte zu berechnende Anwendung auf eine bestimmte Rechenhierarchie abgebildet werden könne (Abs. [0004]).

97 Im Streitpatent ist keine bestimmte Aufgabe formuliert. Vor dem Hintergrund des dargestellten Standes der Technik lässt sich das technische Problem jedoch dahingehend beschreiben, ein verbessertes Verfahren zum Betrieb eines heterogenes Rechensystem und ein ebensolches heterogenes Rechensystem bereitzustellen, um eine Rechenaufgabe effizienter berechnen zu können (Abs. [0017]).

98 Zur Lösung dieses Problems schlägt das Streitpatent ein Verfahren mit den Merkmalen des Patentanspruchs 1 vor, die sich wie folgt untergliedern lassen:

1. Verfahren zum Betreiben eines heterogenen Rechensystems (10), das eine Vielzahl von Rechenknoten (20) und eine Vielzahl von Booster-Knoten (22) aufweist;
2. mindestens einer der Vielzahl von Rechenknoten (20) und der Vielzahl von Booster-Knoten (22) ist so eingerichtet, dass er eine Rechenaufgabe berechnet, wobei die Rechenaufgabe eine Vielzahl von Unteraufgaben aufweist;

3. in einer ersten Iteration der Berechnung wird die Vielzahl von Unteraufgaben in einer ersten Verteilung einem der Vielzahl von Rechenknoten (20) und Booster-Knoten (22) zugewiesen und von diesen verarbeitet;

4. Informationen,

   4.1 die sich auf die Verarbeitung der Vielzahl von Unteraufgaben durch die Vielzahl von Rechenknoten (20) und Booster-Knoten (22) beziehen,

   4.2 werden verwendet, um eine weitere Verteilung der Unteraufgaben zwischen den Rechenknoten (20) und Booster-Knoten (22) zu erzeugen, um sie in einer weiteren Recheniteration zu verarbeiten.

99 Auf das weiterhin offenbarte Rechensystem mit den Merkmalen nach dem unabhängigen Vorrichtungsanspruch 9 des Streitpatents kommt es im Streitfall nicht an.

II. Auslegung

100 Anspruch 1 des Streitpatents betrifft ein Verfahren zum Betrieb eines heterogenen Rechensystems. Das Rechensystem weist eine Vielzahl von Rechenknoten und Booster-Knoten auf zur Berechnung einer eine Vielzahl von Unteraufgaben aufweisenden Rechenaufgabe. Das Verfahren umfasst im ersten Schritt, dass die Vielzahl von Unteraufgaben den Rechenknoten und Booster-Knoten zugewiesen und von diesen verarbeitet wird, und in einem zweiten Schritt, dass Informationen, die sich auf die Verarbeitung der Unteraufgaben durch die Knoten beziehen, verwendet werden, um eine weitere Verteilung der Unteraufgaben zwischen den Knoten zu erzeugen. Die Einzelheiten des Anspruchs 1 bedürfen der Auslegung, soweit sie für die Entscheidung des Rechtsstreits relevant sind.

1. Auslegungsgrundsätze

101 Gemäß Art. 69 EPÜ und dem Protokoll über dessen Auslegung ist ein Patentanspruch nicht nur der Ausgangspunkt, sondern die entscheidende Grundlage für die Bestimmung des Schutzumfangs eines europäischen Patents. Die Auslegung eines Patentanspruchs hängt nicht allein von der strengen, wörtlichen Bedeutung des verwendeten Wortlauts ab. Vielmehr müssen die Beschreibung und die Zeichnungen stets als Erläuterungshilfen für die Auslegung des Patentanspruchs herangezogen werden und nicht nur zur Klärung etwaiger Unklarheiten im Patentanspruch. Dies bedeutet jedoch nicht, dass der Patentanspruch lediglich als Leitfaden dient und dass sich sein Gegenstand auch auf das erstreckt, was nach Prüfung der Beschreibung und der Zeichnungen als der Gegenstand erscheint, für den der Patentinhaber Schutz begehrt (Berufungsgericht, Anordnung v. 23.02.2023, UPC_CoA_335/2023; Anordnung v. 13.05.2024, UPC_CoA_1/2024; Zentralkammer München, Entscheidung v. 16.07.2024, UPC_CFI_1/2023; Lokalkammer Paris, Entscheidung v. 04.07.2024, UPC_CFI_230/2023; Lokalkammer München, Entscheidung v. 31.07.2024,

UPC_CFI_233/2023; Lokalkammer Hamburg, 26.08.2024, UPC_CFI_54/2023; Lokalkammer Düsseldorf, Entscheidung v. 10.10.2024, UPC_CFI_363/2023; Entscheidung v. 31.10.2024, UPC_CFI_373/2024).

102 Der Patentanspruch ist aus der Sicht einer Fachperson auszulegen. Bei der Auslegung eines Patentanspruchs wendet der Fachperson kein philologisches Verständnis an, sondern bestimmt die technische Bedeutung der verwendeten Begriffe mit Hilfe der Beschreibung und der Zeichnungen. Ein Merkmal in einem Patentanspruch ist immer unter Berücksichtigung des gesamten Anspruchs auszulegen (Berufungsgericht, UPC_CoA_1/2024, Beschluss vom 13. Mai 2024). Aus der Funktion der einzelnen Merkmale im Zusammenhang mit dem gesamten Patentanspruch muss abgeleitet werden, welche technische Funktion diese Merkmale einzeln und insgesamt tatsächlich haben. Aus der Beschreibung und den Zeichnungen kann sich ergeben, dass die Patentschrift Begriffe eigenständig definiert und in dieser Hinsicht ein patenteigenes Lexikon darstellt. Selbst wenn die im Patent verwendeten Begriffe vom allgemeinen Sprachgebrauch abweichen, kann es daher sein, dass letztlich die sich aus der Patentschrift ergebende Bedeutung der Begriffe maßgeblich ist (Zentralkammer München, Entscheidung v. 16.07.2024, UPC_CFI_1/2023; Lokalkammer Düsseldorf, Entscheidung v. 31.10.2024, UPC_CFI_373/2024).

2. Fachperson

103 Die Parteien sind sich darüber einig, dass die vom Klagepatent angesprochene durchschnittliche Fachperson einen Universitätsabschluss als Master of Science oder Diplomingenieur/in der Fachrichtung Elektrotechnik oder Computerwissenschaften hat. Weiterhin geht das Gericht davon aus, dass die Fachperson über mehrjährige Berufserfahrung im Bereich des Hochleistungsrechnens hat. Soweit die Beklagten vorschlagen, dass die Fachperson über Berufserfahrung speziell in der Konzeption und Entwicklung von Lösungen zur Berechnung und Verteilung von Rechenaufgaben in einer heterogenen Rechenumgebung verfügen sollte, erscheint dies zu eng gefasst. Diese spezielle fachliche Erfahrung ist jedoch als Teilaspekt der Berufserfahrung im Bereich des Hochleistungsrechnens umfasst.

3. Rechenaufgabe und Unteraufgaben

104 Das Verfahren nach Anspruch 1 des Streitpatents dient gemäß Merkmal 1 dem Betrieb eines heterogenen Rechensystems, das eine Vielzahl von Rechenknoten und eine Vielzahl von Booster-Knoten aufweist. Gemäß Merkmal 2 des streitgegenständlichen Patentanspruchs 1 soll mindestens einer der Vielzahl von Rechnerknoten und der Vielzahl von Booster-Knoten so eingerichtet sein, dass er eine Rechenaufgabe berechnet, die eine Vielzahl von Unteraufgaben aufweist. Erforderlich ist demnach allein, dass einer der Knoten geeignet ist, eine Rechenaufgabe mit einer Vielzahl von Unteraufgaben zu berechnen.

a)

105 Das Merkmal 2 trifft keine Aussage darüber, ob den Rechenknoten für die Berechnung der Rechenaufgabe Booster-Knoten zugewiesen werden. Noch weniger gibt es einen Hinweis auf eine dynamische Zuweisung von Booster-Knoten zu Rechenknoten, wie sie im Stand der Technik aus der WO‘247 bekannt war (vgl. Abs. [0002]). Dies hat im Anspruch 1 des Streitpatents keinen Niederschlag gefunden. Auch die Merkmale 3 und 4 verhalten sich nur zu einer Zuweisung der Unteraufgaben zu der Vielzahl von Rechenknoten und Booster-Knoten, nicht aber zu einer Zuweisung von Knoten zueinander.

106 In dieser Hinsicht ist zu berücksichtigen, dass es sich bei dem Anspruch 1 des Streitpatents um einen Verfahrensanspruch handelt. Das Merkmal 2 ist so formuliert, dass es Anforderungen an das heterogene Rechensystem mit Rechen- und Booster-Knoten aufstellt, das durch das patentgemäße Verfahren betrieben werden soll. Es muss geeignet sein, eine Rechenaufgabe, die eine Vielzahl von Unteraufgaben aufweist, zu berechnen. Merkmal 2 begründet allenfalls insofern einen Verfahrensschritt, dass mit Blick auf Merkmale 3 und 4 die Rechenaufgabe tatsächlich berechnet wird. Ansonsten enthält aber weder Merkmal 2 noch der übrige Anspruch einen Verfahrensschritt, in dem den Rechenknoten Booster-Knoten zugewiesen werden oder umgekehrt.

107 Das Streitpatent schließt eine solche dynamische Zuweisung von Rechen- und Booster-Knoten aber auch nicht aus. Unteranspruch 4 ist auf eine solche Zuweisung gerichtet, die zudem in dem Ausführungsbeispiel des Streitpatents (Z. 27-29 von Abs. [0018]) beschrieben wird. Anspruch 1 des Streitpatents sieht eine solche Art der Zuweisung jedoch nicht zwingend vor.

b)

108 Das Streitpatent enthält keine abschließende Definition des Begriffs der „Rechenaufgabe“ und ihrer „Unteraufgaben“. Die Fachperson wird eine Rechenaufgabe im Sinne des Streitpatents dahingehend verstehen, dass sie eine für einen Prozessor geeignete Arbeits- oder Rechenanweisung im Sinne von Programmcode enthält, die sich in Unteraufgaben aufteilen lässt. Sie unterscheidet sich insofern von einem bloßen Datensatz, der in eine auf einem Prozessor vorhandene (oder anderweitig ladbare) Anwendung eingegeben wird.

109 Stattdessen ist die Rechenaufgabe strukturell mit der die Daten verarbeitenden Anwendung zu vergleichen, auch wenn sie typischerweise komplexer ist, wenn sie auf einem heterogenen Clustercomputer-System verarbeitet werden soll. Das Streitpatent selbst verwendet jedenfalls im Zusammenhang mit der Darstellung des Standes der Technik ebenfalls den Begriff der Anwendung („application“), aus deren Code Aufgaben („tasks“) abgeleitet werden können, indem der Code mit Markierungen versehen wird („taskified“), um Zusammenhänge zwischen einzelnen Aufgaben kenntlich zu machen und skalierbare Aufgaben hervorzuheben (Abs. [0003]).

110 Als Beispiel für eine Rechenaufgabe nennt das Streitpatent eine „Monte Carlo“-Simulation, bei der ein Effekt unter Verwendung einer Zufallszahl modelliert wird, wobei die Berechnungen viele Male hintereinander wiederholt werden (Abs. [0015]). Die Rechenaufgabe besteht hier also aus einem Algorithmus, der unter Verwendung verschiedener Zufallszahlen unzählige Male angewendet wird.

c)

111 Soweit die Parteien darüber streiten, ob eine Rechenaufgabe zwingend heterogen sein müsse, also in Unteraufgaben aufteilbar sein müsse, die sich für die Lösung durch einen Rechenknoten oder einen Booster-Knoten unterschiedlich eignen, oder auch aus nur homogenen (gleichförmigen) Unteraufgaben bestehen dürfe, ist zu berücksichtigen, dass Gegenstand des Anspruchs nicht die Rechenaufgabe ist, sondern ein Verfahren zum Betreiben eines heterogenen Rechensystems. Das Verfahren und das Rechensystem determinieren für sich genommen nicht die Rechenaufgabe, solange es sich tatsächlich um eine Aufgabe und nicht nur um einen Datensatz oder dergleichen handelt. Es ist daher nicht ausgeschlossen, dass mit dem geschützten Verfahren auf einem heterogenen Rechensystem eine Rechenaufgabe berechnet wird, die völlig gleichförmige Unteraufgaben aufweist, die sich nur zur Bearbeitung durch einen der Knotentypen besonders eignen. Allerdings ergibt sich bereits aus dem Begriff des heterogenen Rechensystems, dass aufgrund der qualitativen technischen Unterschiede zwischen Rechenknoten und Booster-Knoten das Verfahren auch in der Lage sein muss, heterogene Rechenaufgaben zu berechnen

4. Iteration

112 Die Vielzahl der Unteraufgaben sind gemäß den Merkmalen 3 und 4 Gegenstand einer ersten und einer weiteren Iteration oder Recheniteration („computing iteration“). Darunter ist die erneute Ausführung der die Unteraufgabe bzw. die Rechenaufgabe bildenden Rechenanweisung zu verstehen.

a)

113 Weder der Patentanspruch 1 noch die Beschreibung des Streitpatents definieren den Begriff der Iteration. Dieser wird vom Streitpatent vorausgesetzt. Nach dem allgemeinen Begriffsverständnis beschreibt der Begriff „Iteration“ die Wiederholung eines Vorgangs, in der Regel um ihn oder sein Ergebnis zu verbessern (vgl. S. 23 der Duplik vom 13. August 2025 der Beklagten unter Verweis auf das Cambridge Dictionary). Konkret in mathematischer Hinsicht beschreibt eine Iteration die wiederholte Anwendung derselben Rechenanweisung.

b)

114 Auf genau diese Bedeutung weist auch die Beschreibung des Streitpatents hin. Es ist zu berücksichtigen, dass das patentgeschützte Verfahren grundsätzlich auf die Berechnung komplexer Rechenaufgaben auf heterogenen Clustercomputern gerichtet ist. Dazu gehören insbesondere Simulationen, wie sie auch das Streitpatent mit der „Monte Carlo“-Simulation

beispielhaft benennt (vgl. Abs. [0015]). Ausdrücklich heißt es dort, dass die Berechnungen viele Male hintereinander wiederholt werden („the calculations being repeated many times in succession“, Z. 51 f. in Abs. [0015]). Es geht also darum, die Rechenanweisungen mit immer neuen Ausgangswerten oder anhand veränderter Datensätze auszuführen und so beispielsweise einen Näherungswert, einen Grenzwert oder eine Entwicklung über die Zeit darzustellen. Dabei können auch die Ergebnisse einzelner Unteraufgaben die Ausgangswerte für die Berechnung anderer Unteraufgaben oder auch derselben Unteraufgabe (Schleife) darstellen.

115 Soweit die Klägerinnen dem in der mündlichen Verhandlung entgegengehalten haben, dass die zitierte Textstelle (Abs. [0015]) den Begriff der Iteration gerade nicht verwende und eine Iteration im Sinne des Streitpatents die Wiederholung der Berechnung gerade nicht voraussetze, kann dem nicht gefolgt werden. Für die Auslegung des Begriffs der Iteration ist es nicht notwendig, dass dieser in der zitierten Textstelle konkret genannt oder gar definiert wird. Es genügt, wenn sich der Fachperson mit Hilfe der Beschreibung und der Zeichnungen die technische Bedeutung des im Patentanspruch verwendeten Begriffs erschließt. Das ist hier der Fall. Die zitierte Textstelle ist die einzige Stelle in der Beschreibung des Streitpatents, die überhaupt ein konkretes Beispiel für eine Rechenaufgabe und ihre Berechnung darstellt. Die Darstellung stimmt mit dem allgemeinen technischen Verständnis der Fachperson vom Begriff der Iteration überein und es ist kein Grund ersichtlich, von diesem Verständnis abzuweichen.

c)

116 Auch bei der gebotenen funktionalen Betrachtung erscheint dieses Verständnis des Begriffs „Iteration“ technisch sinnvoll. Insofern ist zu berücksichtigen, dass es sich beim Anspruch 1 um einen Verfahrensanspruch handelt, der die Rechenaufgabe zum Gegenstand hat. Die Rechenaufgabe mit ihren Unteraufgaben und die Art und Weise ihrer Berechnung sind demzufolge nicht Teil der technischen Lösung des dem Streitpatent zugrundeliegenden Problems sind, sondern sein Ausgangspunkt.

117 Die technische Aufgabe des Streitpatents besteht darin, die Effizienz bei der Berechnung der Rechenaufgabe durch ein heterogenes Rechensystem zu verbessern (Abs. [0017]). Dies gelingt gemäß den Merkmalen 3 und 4 durch eine bestimmte Umverteilung der Unteraufgaben auf die Rechenknoten und Booster-Knoten auf der Grundlage von Informationen über die Verarbeitung der Unteraufgaben in (mindestens) einer früheren Iteration (vgl. auch Abs. [0006], [0017] und [0018]). Die Informationen geben Auskunft darüber, ob andere Rechen- oder Boosterknoten für die Bearbeitung der Unteraufgaben geeigneter sind. Das patentgemäße Verfahren setzt insofern die iterative Berechnung der Unteraufgaben voraus. Denn gerade weil die Unteraufgaben wiederholt berechnet werden, lassen die aus den vergangenen Iterationen gemäß Merkmal 4 gewonnenen Informationen erst eine Aussage über die zukünftige Verarbeitung durch einen Rechen- oder Boosterknoten zu. Die Informationen gemäß Merkmal 4 beziehen sich daher auf die Verarbeitung der Unteraufgaben durch die Vielzahl der Rechen- und Boosterknoten und nicht auf die Unteraufgabe oder den

Knoten als solches (näher dazu siehe unten). Die Verarbeitung der Unteraufgaben durch einen bestimmten Knoten liefert die Information, aus der abgeleitet werden kann, ob die in einer weiteren Iteration wiederholte Berechnung der Unteraufgabe durch denselben oder einen anderen Knoten effizienter ist. Das Streitpatent liefert aber keine Lösung dafür, wie die Berechnung der Rechenaufgabe optimiert werden könnte, wenn die mit den Unteraufgaben festgelegten Rechenanweisungen nicht wiederholt berechnet würden. Es wird nicht erläutert, inwiefern Informationen über die vergangene Verarbeitung von Rechenanweisungen zu einer zukünftigen Verarbeitung anderer (da nicht wiederholter) Rechenanweisungen beitragen könnten.

118 Daher kann auch der Ansicht der Klägerinnen nicht gefolgt werden, wonach bei der gebotenen funktionalen Auslegung kein Bedürfnis bestehe, für eine weitere Verteilung der Unteraufgaben bis zum Abschluss einer ersten „Iteration“ zu warten, insbesondere wenn einzelne Unteraufgaben sehr komplex seien und ihre Bearbeitung wesentlich länger dauere als die von einfacheren Unteraufgaben. In einem solchen Fall sei nicht ausgeschlossen, dass die Berechnung unterbrochen und die erste Verteilung überprüft werde. Iteration sei demnach der Bearbeitungsschritt bzw. die Bearbeitungszeit, nach dem oder nach der eine Bewertung des Status‘ des Systems und der ersten Verteilung sinnvollerweise vollzogen werden könne und aussagekräftige Informationen für eine weitere, grundsätzlich systemweite Verarbeitung bzw. Verteilung vorhanden seien.

119 Für eine solche Auslegung bietet weder der Begriff der Iteration noch das Streitpatent irgendeinen Anhaltspunkt. Die Klägerinnen reduzieren die Iteration auf den Bearbeitungsabschnitt, nach dem eine Bewertung der Aufgabenbearbeitung und Umverteilung der Unteraufgaben sinnvoll erscheint. Das hat allerdings mit der eigentlichen Bedeutung des Begriffs der Iteration im Sinne einer Wiederholung nichts zu tun. Es gibt auch keinen Hinweis darauf, dass die Iteration nach der Lehre des Streitpatents die von den Klägerinnen behauptete Funktion oder gar Bedeutung haben sollte. Nichts deutet im Streitpatent darauf hin, dass eine Iteration in irgendeiner Weise davon abhängig sein soll, wann eine Prüfung der ersten Verteilung und eine weitere Verteilung sinnvoll erscheinen. Im Gegenteil, die Iteration ergibt sich als solche aus der Art der Rechenaufgabe im Sinne einer iterativen Rechenanweisung und erst die wiederholte Anwendung der Rechenanweisung ermöglicht – wie ausgeführt – auf der Grundlage der Informationen über die vergangene Verarbeitung der Unteraufgaben eine effizientere Verteilung auf die Rechen- und Boosterknoten für eine weitere Iteration – hier verstanden als wiederholte Ausführung derselben Rechenanweisung.

d)

120 Es ist weiter zu berücksichtigen, dass die Klägerin zu 2) in ihrer Stellungnahme zur Mitteilung des EPA im Erteilungsverfahren für das Streitpatent noch eine andere Auffassung vertreten hat und von einem zutreffenden Verständnis des Begriffs der Iteration ausgegangen ist (Anlage BP 7). Solche Stellungnahmen eines Patentanmelders im Erteilungsverfahren stellen zwar kein Auslegungsmaterial für die Auslegung des erteilten Patents im Sinne von Art. 69

Abs. 1 EPÜ dar. Ebenso wenig sind solche Äußerungen des Anmelders im Erteilungsverfahren für die Auslegung in einem nachfolgenden Verletzungs- oder Nichtigkeitsverfahren bindend. Sie können jedoch ein Indiz für das fachmännische Verständnis von der Lehre des erteilten Patents im Prioritätszeitpunkt geben (Berufungsgericht, Anordnung vom 20.12.2024, UPC_CoA_402/2024 - Alexion gg. Samsung Bioepis), weil der Anmelder selbst regelmäßig das beste Verständnis von seiner Erfindung und das zugehörige Fachwissen hat.

121 Die Klägerin zu 2) hat in ihrer Stellungnahme zum entgegengehaltenen Stand der Technik ausgeführt, dass eine Rechenaufgabe beispielsweise in einen ersten Teil und einen zweiten Teil aufgeteilt werden könne und der zweite Teil parallel zum ersten Teil oder nach dem ersten Teil berechnet werden könne. Allerdings sei damit keine Vielzahl von Unteraufgaben offenbart, die in mehreren Iterationen berechnet werde (Anlage BP 7, dort zweiter Absatz auf S. 2). Nach der ursprünglich mitgeteilten Auffassung der Klägerin zu 2) geht damit die iterative Berechnung der Rechenaufgabe bzw. ihrer Unteraufgaben über die bloße Aufteilung einer Rechenaufgabe in Unteraufgaben und die parallele oder sukzessive Berechnung dieser Unteraufgaben hinaus. Eine parallele Berechnung scheidet bereits deshalb aus, weil die erste und die weitere Iteration zwingend zeitlich nacheinander erfolgen. Aber auch eine bloß sukzessive Berechnung von Unteraufgaben einer Rechenaufgabe genügt nicht, da es ihr an der wiederholten Ausführung derselben Rechenanweisung fehlt. Eine Iteration ist stattdessen darauf gerichtet, die in der Unteraufgabe enthaltene Rechenanweisung wiederholt – dann mit veränderten Werten – auszuführen, um aus den verschiedenen Ergebnissen der wiederholten Anwendung weitere Erkenntnisse abzuleiten, die infolgedessen über das Resultat der einmaligen Berechnung der Unteraufgabe hinausgehen.

5. <u>Erste und weitere Verteilung der Unteraufgaben</u>

122 Bildet die Rechenaufgabe mit ihren iterativ zu berechnenden Unteraufgaben den Ausgangspunkt des Verfahrens, werden diese Unteraufgaben für die erste Iteration nun in einem ersten Verfahrensschritt gemäß Merkmal 3 einigen der Vielzahl von Rechen- und Boosterknoten zugewiesen und von diesen verarbeitet. In einem zweiten Verfahrensschritt gemäß Merkmal 4 wird eine weitere Verteilung der Unteraufgaben zwischen den Rechen- und Boosterknoten erzeugt.

a)

123 Die weitere Verteilung nach Merkmal 4 muss nicht zwingend alle Unteraufgaben umfassen, die auch Gegenstand der ersten Verteilung waren. Es ist nach der Lehre des Streitpatents zulässig, dass nur ein Teil der ursprünglichen Unteraufgaben umverteilt wird. Auch dies stellt eine weitere Verteilung der Unteraufgabe zwischen den Rechen- und Boosterknoten dar. Ebenso ist es möglich, dass mehrere Unteraufgaben, die von verschiedenen Knoten bearbeitet wurden, in einer weiteren Verteilung von ein- und demselben Knoten berechnet werden.

124 Der Wortlaut der Merkmale 3 und 4 zwingt nicht zu einer Auslegung, nach der alle Unteraufgaben, die bereits Gegenstand der ersten Verteilung waren, auch von der weiteren Verteilung erfasst werden. Dass Merkmal 4 mit der weiteren Verteilung „der Unteraufgaben“ sprachlich auf „die Vielzahl von Unteraufgaben“ gemäß Merkmal 3 rückbezogen ist, bedeutet nicht, dass auch exakt derselbe Satz an Unteraufgaben erneut verteilt werden muss. Sprachlich lässt der Anspruch auch eine Auslegung zu, nach der nur einige der Unteraufgaben erneut verteilt werden, weil auch das eine weitere Verteilung der Unteraufgaben (unter Belassung eines Teils bei den Knoten, denen sie ursprünglich zugewiesen waren) darstellt.

125 Es ist insofern zu berücksichtigen, dass sich während der ersten Recheniteration ergeben kann, dass einzelne Unteraufgaben für die Verarbeitung durch bestimmte Knoten nicht geeignet sind. Es erscheint angebracht, nur diese Unteraufgaben umzuverteilen. Genau dazu dient auch die Verwendung der Informationen bezüglich der Verarbeitung der Vielzahl von Unteraufgaben durch die Vielzahl der Knoten in Merkmal 4. Es erschließt sich nicht, warum auch solche Unteraufgaben, die durch einen geeigneten Knoten verarbeitet werden und für die auch sonst kein Grund für eine Zuweisung zu einem anderen Knoten besteht, umverteilt werden sollten.

126 Genau dies ergibt sich auch aus der Beschreibung des Streitpatents. Demnach können sich bestimmte – mithin nicht alle – Unteraufgaben („certain sub-tasks“), die in einer ersten Verteilung einem Booster zugewiesen wurden, als ungeeignet für die Verarbeitung durch den Booster erweisen, sodass eine Verarbeitung der Teilaufgaben durch einen Rechenknoten anstelle eines Boosters die Berechnung der Aufgabe insgesamt optimieren könnte (Abs. [0017]). Das Streitpatent schlägt dementsprechend eine weitere Iteration der Aufgabe mit einer geänderten weiteren Unteraufgabenverteilung vor, um die Effizienz der Berechnung der Aufgabe zu verbessern (Abs. [0017]).

127 In der Beschreibung des Streitpatents wird darauf hingewiesen, dass die (weitere) Verteilung der Unteraufgaben auch durch die Abhängigkeit von Unteraufgaben untereinander beeinflusst werden kann. Erfordert etwa eine erste Teilaufgabe, die von einem Booster bearbeitet wird, Eingaben von einer zweiten Teilaufgabe, die nicht vom Booster bearbeitet wird, kann dies – so die Patentschrift – zu einer Unterbrechung bei der Bearbeitung der ersten Teilaufgabe führen. Dementsprechend kann nach einer weiteren Verteilung in einer weiteren Iteration sowohl die erste als auch die zweite Teilaufgabe vom Booster bearbeitet werden (Abs. [0022]). Auch hier ist lediglich die Umverteilung einer der Teilaufgaben mit der weiteren Verteilung erforderlich.

128 Nach alledem macht es weder die bessere Eignung eines Rechen- oder Boosterknotens für die Bearbeitung einer Unteraufgabe, noch die damit verbundene Effizienzsteigerung bei der Berechnung der Rechenaufgabe erforderlich, sämtliche Unteraufgaben, die Gegenstand der ersten Verteilung waren, mit der weiteren Verteilung erneut zuzuweisen.

b)

129 Entscheidend ist jedoch, dass das Streitpatent zur Lösung des technischen Problems die weitere Verteilung von Unteraufgaben zwischen den Rechen- und Boosterknoten nach der ersten Iteration vorsieht. Die weitere Verteilung gemäß Merkmal 4 ermöglicht eine Zuweisung von Unteraufgaben an Rechen- oder Boosterknoten, die für die Verarbeitung dieser Unteraufgaben besser geeignet sind, so dass die Unteraufgaben und damit die gesamte Rechenaufgabe effizienter verarbeitet werden können. Dadurch wird das dem Streitpatent zugrundeliegende technische Problem gelöst, den Betrieb eines heterogenen Rechensystems zu verbessern, insbesondere die Verarbeitung von Rechenaufgaben zu optimieren (vgl. Abs. [0017]).

130 Die Lehre von Anspruch 1 des Streitpatents schließt nicht aus, dass einzelne Unteraufgaben in einer der Recheniterationen gelöst und abgeschlossen werden oder sogar inhaltlich neu zugeschnitten werden. Das ist aber nicht Gegenstand der Erfindung und auch nicht die Lösung des technischen Problems. Diese ist stattdessen darauf gerichtet, den von Anfang an vorhandenen und im wesentlichen unveränderten Satz an iterativ zu berechnenden Unteraufgaben einer bestimmten Rechenaufgabe erneut zu verteilen und so eine verbesserte Verarbeitung der Unteraufgabe zu ermöglichen. Dem Streitpatent lässt sich nicht entnehmen, dass die Rechenaufgabe in neue Unteraufgaben aufgeteilt wird, die dann Gegenstand der weiteren Verteilung sind.

6. <u>Informationen, die sich auf die Verarbeitung der Vielzahl von Unteraufgaben beziehen</u>

131 Die weitere Verteilung nach Merkmal 4 soll nach der Lehre des Streitpatents auf der Grundlage der Verwendung von Informationen erzeugt werden, die sich auf die Verarbeitung der Vielzahl von Unteraufgaben durch die Vielzahl von Rechenknoten und Booster-Knoten beziehen. Die Informationen müssen demnach für die weitere Verteilung kausal sein.

a)

132 Wie bereits aus dem Wortlaut des Merkmals 4 ersichtlich ist, muss es sich bei den dort genannten Informationen um solche handeln, die eine Aussage über die Unteraufgaben selbst und ihre tatsächliche Verarbeitung durch die Rechen- und Boosterknoten treffen, wobei die Verarbeitung durch die erste oder jedenfalls vorangehende Recheniterationen bedingt ist. Es genügt demnach nicht, dass die Informationen auf die Rechen- und Boosterknoten als solche bezogen sind und beispielsweise ihre Aktivität, ihr Auslastung, ihre Kapazität oder dergleichen wiedergeben (im Folgenden als „Status“ oder „Status der Knoten“ bezeichnet).

b)

133 Diese Auslegung ergibt sich auch aus der Beschreibung des Streitpatents, die ausdrücklich zwischen Informationen, die sich auf die Verarbeitung der Unteraufgaben beziehen, und dem Status der Knoten unterscheidet. So heißt es zu einem Ausführungsbeispiel:

> “Accordingly, the system 10 includes a mechanism whereby each of the computation nodes and the cluster nodes are arranged such that daemons 26a, 26b and 32 feed back information to the daemon 34 relating to the processing of sub-tasks and a current state of the respective processing entity.“ (Abs. [0018]).

In einer bevorzugten Ausführungsform dieses Ausführungsbeispiels, in dem ein Skalierungsfaktor für jede Unteraufgabe von einem Operator abgeschätzt und in das System eingegeben wurde, heißt es ebenso:

> „During a first iteration of the task, the results of the execution of the sub-tasks are collected together with the information from the daemons concerning the processing of the sub-tasks and the status of the nodes.“ (Abs. [0019])

134 Das Streitpatent unterscheidet demnach ausdrücklich zwischen Informationen in Bezug auf die Verarbeitung der Unteraufgaben und dem Status der Knoten. Beide sollen von den Daemons der Knoten an den Anwendungsmanager mitgeteilt werden, um auf dieser Grundlage über eine erneute Verteilung der Unteraufgaben zu entscheiden. Allerdings haben nur die Informationen, die sich auf die Verarbeitung der Unteraufgaben beziehen, Eingang in Anspruch 1 des Streitpatents gefunden. Dies schließt nicht aus, dass auch der Status der Knoten mitgeteilt und bei der weiteren Verteilung der Unteraufgaben Berücksichtigung findet. Entscheidend ist aber, dass zwischen den Informationen betreffend die Verarbeitung der Unteraufgaben und dem Status der Knoten zu unterscheiden ist und erstere gemäß der Lehre des Streitpatents zwingend für die Erzeugung der weiteren Verteilung verwendet werden müssen.

135 Eine andere Auslegung kann auch nicht damit begründet werden, dass in der Beschreibung weiter ausgeführt wird:

> “For each iteration, the daemons operating in the computation nodes and the boosters report status information to the application manager and the resource manager enabling the calculation of subsequent iterations to be optimized by a further adjustments to the allocation of sub-tasks to computation nodes and boosters.” (Abs. [0020])

Soweit hier von Statusinformationen die Rede ist, die berichtet werden, greift die Textstelle lediglich das zuvor beschriebene Ausführungsbeispiel auf. Unter den Statusinformationen sind die Informationen betreffend die Verarbeitung der Unteraufgaben, d.h. der Verarbeitungsstatus, und der Status der Knoten zu verstehen. Es gibt keinen Anhaltspunkt dafür, dass das Streitpatent hier ausschließlich auf den Status der Knoten Bezug nimmt, der berichtet werden soll, und dass die in Merkmal 4 von Anspruch 1 genannten Informationen auch allein diesen Status der Knoten umfassen können. Dem steht schon der Wortlaut des Merkmals entgegen, das allein Informationen betrifft, die sich auf die Verarbeitung der Vielzahl

von Unteraufgaben beziehen („information relating to the processing.of the plurality of sub-tasks“).

136 Zu den Informationen, die sich auf die Verarbeitung der Unteraufgaben beziehen, können Informationen über die Eignung der Unteraufgaben für die Bearbeitung durch den entsprechenden Rechen- oder Boosterknoten gehören (Abs. [0017] bis [0019]), wie etwa die Skalierbarkeit der Unteraufgabe (vgl. Z. 5-7 von Abs. [0022]). Aber auch Informationen, die sich im Laufe der Verarbeitung einer Unteraufgabe ergeben, etwa weil die Berechnung der einen Unteraufgabe von einer anderen Unteraufgabe abhängig ist, sind als Informationen im Sinne von Merkmal 4 anzusehen, die sich auf die Verarbeitung der Unteraufgaben beziehen und die weitere Verteilung bedingen können. In der Streitpatentschrift ist dies wie folgt für ein Ausführungsbeispiel beschrieben:

> “(…) the distribution may also be influenced by information learned about the processing of the sub-task and any need to call further sub-tasks during the processing. If a first sub-task being handled by a booster requires input from a second sub-task not being handled by the booster, this may lead to an interruption in the processing of the first sub-task. Accordingly, the daemon at the booster handling the first sub-task can report this situation to the application manager such that in a further iteration both the first and second subtasks are handled by the booster.”

137 Die zitierten Ausführungsbeispiele machen auch deutlich, dass die Informationen im Sinne von Merkmal 4 nicht nur allgemein den Status der Verarbeitung der Gesamtheit der Unteraufgaben betreffen, sondern ihren Ausgangspunkt in der Verarbeitung der einzelnen Unteraufgabe durch den Rechen- oder Boosterknoten haben, dem sie zugewiesen ist, und die weitere Verteilung auf Basis dieser Informationen erzeugt wird. Insofern ist es nicht erforderlich, dass sämtliche Unteraufgaben umverteilt werden, sondern nur solche, für die sich die Notwendigkeit aus der jeweiligen Information ergibt.

c)

138 Gegen diese Auslegung kann nicht mit Erfolg eingewandt werden, die in Merkmal 4 von Anspruch 1 des Streitpatents genannten Informationen umfassten auch Statusinformationen der Knoten, weil das Streitpatent gegenüber dem Stand der Technik erkannt habe, dass das Verarbeiten der Unteraufgaben weitergehenden Einfluss auf den Status der Knoten habe und nicht nur deren Auslastung betreffe, wenn beispielsweise auch die Eignung des Knotens für die Unteraufgabe in den Blick genommen werde. Schon diese Argumentation zeigt, dass die Information zwingend auf der Verarbeitung der Unteraufgabe als solche beruhen muss, damit sich die Lehre des Streitpatent überhaupt von dem – unstreitig – im Stand der Technik bekannten Lastenausgleich zwischen den Knoten auf Basis ihrer Statusinformationen abgrenzen kann. Wenn der Erfindungsgedanke gerade darin liegt, die Information betreffend die Verarbeitung der Unteraufgaben zu nutzen, muss dies auch Gegenstand des Anspruchs sein. Genau dies ergibt sich so auch aus dem Wortlaut von Merkmal 4.

139 Mit dieser Begründung ist auch den Folgerungen der Klägerinnen aus der Argumentation entgegenzutreten, die dem Streitpatent zugrundeliegende Erkenntnis bestehe darin, die im Stand der Technik vorbekannte Berücksichtigung der Lastenverteilung erstmals in den Zusammenhang der tatsächlichen Verarbeitung der Unteraufgaben zu stellen. Selbst wenn dem so ist, kann daraus nicht gefolgert werden, die Lastenverteilung oder der sonstige Status eines Knotens seien als Information, die sich auf die Verarbeitung der Unteraufgabe bezieht, ausreichend. Denn dies ist keine Information, die spezifisch die Verarbeitung der jeweils zugewiesenen Unteraufgabe betrifft.

140 Anders kann auch die Einlassung der Klägerin zu 2) im Erteilungsverfahren nicht verstanden werden, die ein weiteres Indiz für die Auslegung darstellt. Die Klägerin zu 2) nimmt darin zur Erläuterung ihres Verständnisses von den Informationen im Sinne von Merkmal 4 auf genau die zuvor zitierten Ausführungsbeispiele des Streitpatents Bezug und beschränkt die Informationen in Abgrenzung zum Stand der Technik auf solche über die Skalierbarkeit der Unteraufgabe und solche, die aus der Verarbeitung der Unteraufgabe und der Notwendigkeit, während der Verarbeitung weitere Unteraufgaben aufzurufen, gewonnen wurden. Die Klägerin zu 2) erklärt explizit, dass sich die Erfindung nach dem Streitpatent auf die Analyse der Unteraufgaben und den Erhalt von Informationen bezieht, die auf die Verarbeitung der Vielzahl der Unteraufgaben bezogen sind, und nicht auf die bloße Lastverteilung (Anlage BP 7, dort dritter Absatz auf S. 2).

d)

141 Sofern die Klägerinnen erstmals in der mündlichen Verhandlung die Auffassung vertreten haben, auch Merkmale oder Eigenschaften einer Rechenaufgabe, die vom Nutzer vorab festgelegt worden seien, seien als Informationen im Sinne von Merkmal 4 anzusehen, kann dem nicht gefolgt werden.

142 Nach dem Wortlaut des Anspruchs müssen sich die Informationen gemäß Merkmal 4 auf die Verarbeitung der Vielzahl von Unteraufgaben durch die Vielzahl von Rechen- und Boosterknoten beziehen und nicht nur auf die Rechenaufgabe oder die Unteraufgaben als solches. Die Lehre des Streitpatents besteht gerade darin, aus der Verarbeitung der Unteraufgaben in vorangehenden Iterationen Informationen zu gewinnen, die es ermöglichen zu erkennen, dass die erneute Ausführung der Unteraufgaben auf anderen Rechen- und Boosterknoten effizienter möglich ist, und die Unteraufgaben entsprechend umzuverteilen. Vom Anwender im Voraus festgesetzte Merkmale oder Eigenschaften der Rechenaufgabe oder ihrer Unteraufgaben stehen dazu im Widerspruch, weil sie über die tatsächliche Verarbeitung durch den Rechen- oder Boosterknoten, dem die Unteraufgabe in der ersten Iteration zugewiesen ist, keine Aussage treffen. Es gibt keine Information über die Verarbeitung der Unteraufgabe, die für die weitere Verteilung der Unteraufgabe kausal werden könnte.

143 Etwas anderes ergibt sich auch nicht aus der Beschreibung des Streitpatents, wonach ein Nutzer im Vorfeld der Aufgabenbearbeitung für jede Unteraufgabe einen Skalierbarkeitsfaktor

abschätzt, der in das Rechensystem eingegeben wird und Grundlage für die erste Verteilung der Unteraufgabe ist (Abs. [0009] und [0019]; vgl. auch Abs. [0003] zum Stand der Technik). Diese vom Nutzer angegebenen Informationen haben lediglich Einfluss auf die erste Verteilung und die erste Iteration (vgl. „initial distribution“ in Abs. [0009]). Die weitere Verteilung, um die es in Merkmalsgruppe 4 allein geht, basiert hingegen ausschließlich auf Informationen, die aus der Verarbeitung der Unteraufgaben in den vorhergehenden Iterationen gewonnen wurden (vgl. „During a first iteration of the task, the results (…) are collected together with the information from the daemons concerning the processing of the sub-tasks (…). The application manager then performs a reassignment of the sub-tasks (…)“ in Abs. [0019]). Die vom Anwender anfänglich bereitgestellten Informationen spielen demnach für die weitere Verteilung keine Rolle.

III. Verletzung

144 Die angegriffene Ausführungsform ist unabhängig davon, ob sie mit oder ohne vorinstallierter Run:ai-Software angeboten und geliefert wird, nicht geeignet, das mit dem Anspruch 1 des Streitpatents geschützte Verfahren anzuwenden.

1. Angegriffene Ausführungsform

145 Mit der Verletzungsklage angegriffen sind die von den Beklagten beworbenen und in Verkehr gebrachten DGX-System-Produkte unabhängig davon, ob sie mit oder ohne vorinstallierter Run:ai-Software vertrieben werden.

146 Bei Nvidia DGX handelt es sich um eine Reihe von Nvidia-Servern und -Workstations, die neben der Verwendung von CPUs schwerpunktmäßig auf die Verwendung von Grafikprozessoren ('GPUs') zur Beschleunigung von Deep-Learning-Anwendungen spezialisiert sind. Gemäß der (nicht abschließenden) Aufzählung der Klägerinnen gehören zu diesen DGX-Systemen solche Systeme und Cluster wie DGX H100, DGX A100, DGX H200, DGX GH200, DGX B200 oder DGX GB200 SuperPOD. Alle diese Systeme sind Vertreter der angegriffenen Ausführungsform, weil sie nach dem unbestrittenen Vortrag der Beklagten hinsichtlich des Verletzungsvorwurfs identisch sind.

147 Die Klägerinnen haben in der Klageschrift erklärt, bei der angegriffenen Ausführungsform handele es sich um von den Beklagten angebotene bzw. vertriebene DGX-System-Produkte, die mit Run:ai-Software geliefert werden. Nachdem die Beklagten jedoch in der Klageerwiderung bestritten haben, dass die Run:ai-Software beim Kauf eines DGX-Produkts noch nicht installiert sei, sondern vom Kunden eigenständig heruntergeladen und installiert werden müsse, haben die Klägerinnen in der Replik erklärt, an einer mittelbaren Verletzung könne kein Zweifel bestehen, selbst wenn die Run:ai-Software nicht im Bündel mit den DGX-Systemen angeboten und geliefert werde. Entgegen der Auffassung der Beklagten ist mit diesem Vortrag der Klägerinnen keine Klageänderung oder -erweiterung im Sinne von Regel

263 VerfO verbunden, die eines besonderen Antrags und einer gesonderten Zulassung bedarf.

a)

148 Nach der Rechtsprechung des Berufungsgerichts stellt nicht jedes neue Argument eine Klageänderung dar, die eine Partei dazu verpflichtet, eine Zulassung gemäß Regel 263 RoP zu beantragen. Eine Klageänderung tritt ein, wenn sich die Art oder der Umfang der Streitigkeit ändert. In einem Verletzungsverfahren ist dies beispielsweise der Fall, wenn der Kläger ein anderes Patent geltend macht oder ein anderes Produkt beanstandet (Berufungsgericht, Anordnung v. 21.11.2024, UPC_CoA_456/2024 – OrthoApnea).

149 Für die Art und den Umfang eines Rechtsstreits sind einerseits die gestellten Anträge maßgeblich, aber auch die Tatsachen, die zur Begründung dieser Anträge und der zugrundeliegenden Ansprüche vorgetragen werden. Mit dem Berufungsgericht ist davon auszugehen, dass nicht jeder weitere oder geänderte Tatsachenvortrag zugleich zu einer Klageänderung führt. Dies gilt auch dann, wenn sich das angegriffene Produkt nach der Klageerwiderung des Beklagten im Detail anders darstellt als es der Kläger in der Klageschrift beschrieben hatte, solange sich der Verletzungsangriff weiterhin tatsächlich auf das allgemein beschriebene Produkt bezieht. Nicht immer sind dem Kläger sämtliche technischen Einzelheiten eines angegriffenen Produkts bekannt, so dass es ihm möglich sein muss, auf Klarstellungen und Berichtigungen des Beklagten in der Klageerwiderung zu reagieren und seinen Vortrag insofern anzupassen.

b)

150 So liegt der Fall auch hier. Soweit die Klägerinnen nunmehr in Kauf nehmen, dass die Run:ai-Software gegebenenfalls auch erst von den Kunden installiert wird, und ihre Klage auch auf diese Sachverhaltskonstellation stützen, beanstanden sie kein gänzlich anderes Produkt. Sie greifen weiterhin die DGX-System-Produkte an und stellen im Rahmen der mittelbaren Verletzung weiterhin auf die Verwendung der Run:ai-Software ab mit dem einzigen Unterschied, dass diese nicht schon bei der Lieferung installiert ist, sondern erst von den Abnehmern installiert wird. Art und Umfang des Rechtsstreits sind auch mit der Replik der Klägerinnen im Wesentlichen identisch geblieben.

151 Der Klageantrag der Klägerinnen knüpft an den Vorwurf einer mittelbaren Verletzung durch die Beklagten an. In ihrer Replik verfolgen die Klägerinnen identische Klageanträge und Ansprüche weiter. Soweit die Klage auch auf eine unmittelbare Patentverletzung gestützt werden sollte, wurde diese nicht zugelassen.

152 Auch die den Vorwurf der mittelbaren Verletzung begründende Tatsachengrundlage hat sich nicht derart wesentlich geändert, dass dies eine Klageänderung gemäß Regel 263 VerfO begründen könnte. Der Tatsachenvortrag zu den Verletzungshandlungen im Sinne von Art. 26 Abs. 1 EPGÜ – das Anbieten und Liefern von DGX-Systemen durch die Beklagten – ist

unverändert, da er unabhängig davon ist, ob die Software vorinstalliert ist oder nicht. Auch die gemäß Art. 26 Abs. 1 EPGÜ erforderliche Eignung der DGX-Systeme für die Benutzung des Verfahrens durch die Kunden ist unabhängig davon, ob die Software vorinstalliert ist oder erst von den Kunden installiert wird. Denn die Behauptung der Klägerinnen geht dahin, dass das geschützte Verfahren mit den DGX-Systemen einschließlich der darauf installierten Run:ai-Software angewendet werden kann. Die Eignung der DGX-Systeme zur Softwareinstallation ist aber, unabhängig davon wo und wann sie erfolgt, unstreitig. Insofern stellt sich auch die Frage, ob es sich bei den DGX-Systemen ohne vorinstallierte Software um ein Mittel, dass sich auf ein wesentliches Element der Erfindung bezieht, nicht gänzlich anders als wenn die Software vorinstalliert ist. Dies wird von keiner der Parteien erörtert. Der Zeitpunkt der Installation hat nach alledem allenfalls Einfluss darauf, ob die Beklagten im Sinne von Art. 26 Abs. 1 EPGÜ wussten oder hätten wissen müssen, dass die DGX-Systeme – nunmehr ohne vorinstallierte Software – dazu geeignet und bestimmt sind, das patentgeschützte Verfahren anzuwenden. Diese Frage mag sich etwas anders stellen, wenn die DGX-System-Produkte nicht mit vorinstallierter Run:ai-Software angeboten und geliefert werden, sondern die Software erst vom Kunden installiert werden muss. Da es sich aber nur um eine von mehreren Voraussetzungen der mittelbaren Verletzung handelt, deren Beurteilung zudem von weiteren Umständen des Einzelfalls abhängig ist, die unverändert geblieben sind, ist der in der Replik angepasste Vortrag hinsichtlich des Zeitpunkts der Installation der Software, nicht geeignet, die Natur oder den Umfang des Rechtsstreits zu ändern.

2. Eignung zur Anwendung des geschützten Verfahrens bei vorinstallierter Software

153 Die angegriffene Ausführungsform ist für die Anwendung des mit dem Anspruch 1 des Streitpatents geschützten Verfahrens nicht geeignet im Sinne von Art. 26 Abs. 1 EPGÜ. Die Verwirklichung von Merkmal 4 lässt sich nicht feststellen.

a) Rechenaufgabe und Unteraufgaben

154 Es kann dahinstehen, ob die angegriffene Ausführungsform als heterogenes Rechensystem angesehen werden kann, deren CPUs und GPUs eine Vielzahl von Rechenknoten und eine Vielzahl von Boosterknoten im Sinne von Merkmal 1 des Patentanspruchs 1 darstellen. Zwischen den Parteien ist jedenfalls unstreitig, dass ein „Workload“ im Sinne der Terminologie der Run:ai-Software-Dokumentation als Rechenaufgabe und „Pods“ im Sinne dieser Terminologie als Unteraufgaben im Sinne von Merkmal 2 angesehen werden können. Unstreitig ist auch, dass der Run:ai-Scheduler die Pods nach vorbestimmten Regeln den verschiedenen Ressourcen – CPUs und GPUs, die für die weitere Verletzungsprüfung jeweils als Rechen- und Boosterknoten im Sinne des Streitpatents unterstellt werden können – zuweist.

b) Informationen, die sich auf die Verarbeitung der Vielzahl von Unteraufgaben beziehen, und weitere Verteilung

155 Es lässt sich aber nicht feststellen, dass die angegriffene Ausführungsform mit installierter Run:ai-Software Informationen, die sich auf die Verarbeitung der Vielzahl von Unteraufgaben durch die Vielzahl von Rechenknoten und Boosterknoten beziehen, verwendet, um Unteraufgaben gemäß Merkmal 4 erneut zu verteilen. [redacted].

aa)

156 Die Klägerinnen berufen sich für ihre Behauptung auf den Run:ai Scheduler und seine Fähigkeit, für jedes Projekt die Ressourcenzuteilung ständig neu zu berechnen und im laufenden Prozess neu zuzuweisen.

157 Die Run:ai-Software setzt voraus, dass die Organisation oder Geschäftseinheit jeweils in Abteilungen („departments“) und Projekten („projects“) unterteilt wird und die einzelnen Nutzer diesen Abteilungen und Projekten zugewiesen werden. Workloads werden auf der Ebene der Projekte dem Rechensystem übermittelt. Weiterhin setzt die Software voraus, dass die vorhandenen Ressourcen in so genannten Node Pools („node pools“) gruppiert werden. Node Pools sind eine logische Gruppierung von zwei oder mehr „nodes“ – nach der Terminologie von Run:ai handelt es sich um Server – auf der Grundlage bestimmter Merkmale, beispielsweise nach Hardwaretyp. So kann ein Node Pool ausschließlich aus CPU-Servern bestehen oder alle leistungsfähigen GPU-Server eines Clusters umfassen. Schließlich müssen den Projekten feste Kontingente an bestimmten Node Pools zugewiesen werden („reserved quota“). Beispielsweise bedeutet die Zuweisung eines GPU-Kontingents zu einem Projekt auf Basis eines Node Pools, dass die von diesem Projekt eingereichten Workloads berechtigt sind, eine bestimmte Anzahl GPUs als garantierte Ressourcen zu nutzen, und sie für alle Arten von Workloads dieses Projekts verwenden können. Es ist allerdings nicht zwingend, dass die Workloads immer alle garantierten Ressourcen des Node Pools eines Projekts benötigen. Die Workloads können auch weniger Ressourcen erfordern (vgl. Anlagen K 40 und K 47).

158 Grundsätzlich kann ein Workload nur bearbeitet werden, wenn dafür auch die erforderlichen Ressourcen an CPUs und GPUs zur Verfügung stehen (zur Ausnahme durch das „elasticity feature“ siehe unten). Die Pods eines Workloads werden dann auf diese Knoten verteilt. Der Run:ai Scheduler kann einem Workload aber auch mehr als die von ihm geforderten Ressourcen und auch mehr Ressourcen als das garantierte Kontingent des zugehörigen Projekts zuweisen, wenn solche weiteren Ressourcen verfügbar sind. Hat ein Projekt beispielsweise eine „reserved quota“ von vier GPUs, kann der Run:ai-Scheduler einem Workload dieses Projekts auch acht GPUs zuweisen, wenn diese nicht anderweitig belegt sind, mithin vier GPU mehr als erforderlich („over quota“).

bb)

159 Die Run:ai-Software kann die zugewiesenen Ressourcen nun dynamisch umverteilen. Die Klägerinnen berufen sich insofern zunächst auf das „elasticity feature“. Ist dieses aktiviert und stehen für einen zweiten Workload nicht die angeforderten Ressourcen zur Verfügung, kann der Run:ai-Scheduler in seinem nächsten Planungszyklus Ressourcen, die das erforderliche Kontingent des ersten Workloads übersteigen, diesem ersten Workload entziehen und dem zweiten Workload zuweisen. In dem zuvor geschilderten Beispiel hatte ein erster Workload vier GPUs angefordert und der Scheduler hatte ihm acht GPUs zugewiesen, weil vier weitere GPUs verfügbar waren. Erfordert nun ein weiterer Workload ein garantiertes Kontingent von drei GPUs aus demselben Node Pool und sind keine weiteren GPUs vorhanden, kann der Run:ai-Scheduler drei GPUs, die dem ersten Workload zugewiesen waren, dem zweiten Workload zuweisen. Umgekehrt können einem Workload in einem weiteren Planungszyklus auch weitere Ressourcen zugewiesen werden, sollten solche frei werden.

160 Es lässt sich jedoch nicht feststellen, dass die angegriffene Ausführungsform im Zuge der dynamischen Umverteilung Merkmal 4 von Patentanspruch 1 verwirklicht. Die Zuweisung von Ressourcen in dem weiteren Planungszyklus wird nicht unter Verwendung von Informationen erzeugt, die sich auf die Verarbeitung der Vielzahl von Unteraufgaben – hier der Pods des jeweiligen Workloads – durch die Vielzahl von Rechen- und Boosterknoten beziehen. Stattdessen ist allein die Verfügbarkeit der Knoten für die Zuweisung von Pods zu den Knoten maßgeblich. Dies hat mit der Verarbeitung der Unteraufgaben in einer vorhergehenden Iteration nichts zu tun. Eine Umverteilung von Ressourcen findet bereits dann statt, wenn ein Workload über die erforderlichen Ressourcen hinaus Rechenleistung nutzt und wenn ein anderer Workload oder ein anderes Projekt Bedarf an Ressourcen anmeldet, die anderweitig nicht zu Verfügung stehen. Dafür ist der Run:ai-Scheduler überhaupt nicht auf Informationen über die Berechnung der Pods angewiesen. Vielmehr würde ein internes Schema über die jeweils an die in Bearbeitung befindlichen Workloads und Projekte vergebenen Ressourcen genügen, das dem Scheduler in jedem Planungszyklus die derzeit freien oder abziehbaren Ressourcen aufzeigt. Ähnliches gilt, wenn ein Projekt oder Workload beendet ist und Ressourcen frei werden. Ihre Vergabe an einen laufenden Workload steht in keinem Zusammenhang mit der Bearbeitung der Pods dieses Workloads.

cc)

161 Die Klägerinnen stellen für ihre Verletzungsbehauptung weiterhin auf den „fairshare“ und das „fairshare balancing“ ab, das der Run:ai-Scheduler anwenden kann.

162 Ausweislich der Run:ai-Dokumentation können Projekte einen über das garantierte Kontingent an Ressourcen („reserved quota“) hinausgehenden Anteil an ungenutzten Ressourcen eines Node Pools beanspruchen, die so genannten überkontingentierten Ressourcen („over quota“). Der „fairshare“ ist begrifflich die Summe der garantierten Ressourcen des Projekts zuzüglich des Anteils der nicht garantierten Ressourcen in diesem Node Pool („reserved quota“ + „over quota“), der vom Run:ai-Scheduler berechnet wird.

„Fairshare balancing“ bedeutet nun, dass der Run:ai-Scheduler bestrebt ist, jedem Projekt die Ressourcen zur Verfügung zu stellen, die ihm zustehen, wobei er zwei Hauptparameter verwendet, nämlich die verdiente Quote und den verdienten Fairshare (vgl. Anlage K 40).

163 Demnach kann auch im Zuge des „fairshare balancing“ eine Ressourcen-Umverteilung stattfinden. Wenn die Ressourcen eines Node Pools für ein Projekt unter dem „fairshare“ liegen und die eines anderen Projekts über dem „fairshare“, verschiebt der Run:ai-Scheduler Ressourcen zwischen den Projekten. Aber auch in diesem Fall wird das Merkmal 4 von Anspruch 1 des Streitpatents nicht verwirklicht. Es gibt keine Anhaltspunkte, dass diese Umverteilung im Rahmen des „fairshare balancing“ auf der Basis von Informationen über die Verarbeitung der Pods der jeweiligen Workloads erfolgt. Vielmehr ist allein der Wert der reservierten Ressourcen und der überkontingentierten Ressourcen und ihre Verfügbarkeit maßgeblich. Es geht nicht um die Umverteilung von Unteraufgaben eines Workloads, sondern von Ressourcen zwischen verschiedenen Projekten. Die Bearbeitung der einzelnen Pods eines Workloads ist dafür unerheblich.

dd)

164 Eine Verwirklichung von Merkmal 4 lässt sich auch nicht speziell für die „preemptable workloads“ feststellen (vgl. u.a. S. 1 ff. der Anlage K 40a), auf die die Klägerinnen insbesondere in der mündlichen Verhandlung abgestellt haben. Die „preemptability“ eines Workloads ist die Voraussetzung dafür, dass der Workload eine „over quota“ in Anspruch nehmen kann. Umgekehrt können diesen Workloads die Ressourcen, die über die „deserved quota“ hinaus zugewiesen wurden, d.h. die „over quota“, auch wieder entzogen werden, wenn vorrangig zu bearbeitende Workloads diese Ressourcen benötigen und nicht genügend Ressourcen für alle Workloads vorhanden sind. Für diese Entscheidung, ob einem Workload Ressourcen entzogen werden können, benötigt der Run:ai-Scheduler nach Auffassung der Klägerinnen sowohl die Information, dass es sich um einen „preemptable workload“ handele, als auch die Information, dass der Workload Ressourcen „over quota“ nutzt. Diese Informationen stellen nach Auffassung der Klägerinnen Informationen im Sinne von Merkmal 4 dar. Dem vermag sich das Gericht nicht anzuschließen.

165 Die „preemptability“ eines Workloads stellt keine Information dar, die sich auf die Verarbeitung der Unteraufgaben durch die Rechen- und Boosterknoten bezieht, wie dies von Merkmal 4 verlangt wird. Ob ein Workload „preemptable“ ist, wird vom Nutzer festgelegt und stellt eine Eigenschaft des Workloads dar, die festliegt, bevor überhaupt mit der Verarbeitung der einzelnen Pods begonnen wird. Allein der Umstand, dass der Workload derzeit in Bearbeitung ist und einzelne seiner Pods aufgrund der Eigenschaft „preemptable“ von der Verarbeitung ausgenommen werden können, wenn andere Departments, Projekte oder Workloads Ressourcenbedarf haben, macht die Eigenschaft nicht zu einer Information im Sinne von Merkmal 4. Die Information über die „preemptability“ gewinnt ihre Bedeutung erst durch den Umstand, dass es auch Workloads gibt, die keine „over quota“ erhalten und auch nicht unterbrochen werden können, mithin nicht „preemptable“ sind. Damit sind weder die Verarbeitung der Pods durch die CPUs und/oder GPUs noch die Information über die

„preemptability“ kausal für die Unterbrechung der Verarbeitung der Unteraufgaben, sondern allein der Ressourcenbedarf anderer Workloads.

166 Was die vermeintliche Information angeht, welche Pods des Workloads Teil der „over quota“ sind und deren Verarbeitung infolgedessen unterbrochen werden kann, lässt sich eine Verwirklichung von Merkmal 4 ebenfalls nicht feststellen. Es bleibt schon im Dunkeln, auf welcher Ebene überhaupt festgelegt wird, welche Pods eines Workloads durch die als „over quota“ zugewiesenen Ressourcen verarbeitet werden und welche im Rahmen der „deserved quota“. Es erscheint nicht einmal ausgeschlossen, dass sich die Information darauf beschränkt, welche konkreten Ressourcen einem Workload als „over quota“ zugewiesen wurden. Auf den spezifischen Pod und seine Verarbeitung durch diese Ressourcen kommt es dann gar nicht an, so dass es sich nicht um Informationen im Sinne von Merkmal 4 handelt. Es gibt jedenfalls keinen Anhalt dafür, dass ein Pod, für den nicht bereits bei Zuweisung der Ressourcen klar war, dass er „over quota“ bearbeitet wird, ad hoc auf der Grundlage seiner bisherigen Verarbeitung aus mehreren Pods ausgewählt wird, um seine Verarbeitung wegen anderweitigen Ressourcenbedarfs zu beenden. Wie die Klägerinnen selbst in der mündlichen Verhandlung erklärt haben, handelt es sich allenfalls um eine Eigenschaft der Aufgabe, mithin des Pods, wenn gefragt wird, was für eine Aufgabe mit welcher Charakterisierung gerade auf dieser oder jener GPU stattfinde. Die Eigenschaft einer Unteraufgabe stellt jedoch keine Information dar, die sich auf ihre tatsächliche Verarbeitung durch die Rechen- und Boosterknoten bezieht.

167 Ungeachtet dessen ist außerdem festzuhalten, dass der Entzug von Ressourcen nicht zu einer Umverteilung von Pods führt, sondern die Verarbeitung von Pods, die „over quota“ bearbeitet wurden, schlicht beendet wird. Soweit sich die Klägerinnen darauf berufen, dass die Verarbeitung dieser Pods zu einem späteren Zeitpunkt – eventuell auch nach mehreren Planungszyklen – fortgesetzt wird, wenn wieder ausreichend Ressourcen zur Verfügung stehen, führt dies nicht zu einer Verteilung im Sinne von Merkmal 4. Die Informationen, die zur Unterbrechung der Verarbeitung dieser Pods führten, sind nicht kausal für die Zuweisung der Ressourcen. Diese Zuweisung ist nicht durch Informationen bedingt, die sich auf die frühere Verarbeitung der Pods durch die jeweilige CPU oder GPU beziehen. Ursache ist vielmehr allein der Umstand, dass weitere Ressourcen verfügbar sind. Das Konzept, dass der angegriffenen Ausführungsform zugrunde liegt, ist insofern ein gänzlich anderes, als die Lehre des Streitpatents vorsieht.

168 Nach der patentgeschützten technischen Lehre werden Informationen, die sich auf die Verarbeitung der Unteraufgaben durch die Knoten und damit notwendigerweise auf die Verarbeitung in früheren Iterationen beziehen, verwendet, um eine weitere Verteilung zu erzeugen. Die Run:ai-Software zielt hingegen auf die gerechte Verteilung knapper Ressourcen zwischen verschiedenen Nutzern unter vollständiger Auslastung der vorhandenen Ressourcen ab. Für Letzteres sind allenfalls Informationen darüber notwendig, ob eine CPU/GPU verfügbar ist oder gemacht werden kann. Dass es damit zwangsläufig auch auf die Information darüber ankommen kann, dass auf einer CPU/GPU überhaupt eine

Unteraufgabe bearbeitet wird und zu welcher Art von Workload (elastic/non-elastic, preemptable/non-preemptable) sie gehört, macht diese Information jedoch noch nicht zu einer solchen im Sinne von Merkmal 4.

169 Daher führt auch der Verweis Klägerinnen auf die Möglichkeit des „bin-packing“ nicht weiter. In dieser Variante der Software-Anwendung ist der Run:ai-Scheduler bestrebt, eine fragmentierte Belastung der CPUs und GPUs zu vermeiden und Pods dergestalt zwischen den Prozessoren zu verschieben, dass ein Node möglichst vollständig ausgelastet ist und ausreichend GPUs frei sind, um den nächsten Workload in der Warteschlange zu bearbeiten (vgl. S. 4 der Anlage K 40a unter „Bin-packing & Consolidation“). Dass dafür Informationen, die sich auf die Verarbeitung der Pods durch die jeweiligen CPUs und GPUs beziehen, benötigt werden, lässt sich – wie zuvor bereits ausgeführt – nicht feststellen.

ee)

170 Auch die Priorität eines Workloads stellt keine Information im Sinne von Merkmal 4 dar. Prioritäten werden von den Nutzern vergeben. Sie haben zur Folge, dass Workloads mit höherer Priorität vor Workloads mit niedrigerer Priorität zur Bearbeitung eingeplant werden. Es können sogar Workloads mit höherer Priorität solche mit niedrigerer Priorität innerhalb desselben Projekts verdrängen (siehe Anlage K 40). Werden infolgedessen Ressourcen dem einen Workload entzogen und dem anderen zugewiesen, hat dies nichts mit der Bearbeitung der Pods des jeweiligen Workloads zu tun, sondern beruht allein auf der davon unabhängigen Prioritätsinformation, die nicht an den Pods, sondern dem Workload anknüpft.

ff)

171 Schließlich wird das Merkmal 4 von Anspruch 1 des Streitpatents auch nicht durch Inferenz Workloads und die Verwendung des „Hugging Face“ Inferenz Modells verwirklicht.

172 Die Klägerinnen haben ihren Verletzungsvorwurf erstmals in der Replik vom 13. Juni 2025 auf dieses Feature der Run:ai-Software gestützt. Ob dieser Vortrag gemäß Regel 9.2 VerfO unberücksichtigt bleiben kann, kann letztlich dahinstehen. Selbst wenn er berücksichtigt wird, ist damit die Benutzung der Lehre des Streitpatents nicht schlüssig vorgetragen.

173 Es ist schon nicht vorgetragen und auch sonst nicht ersichtlich (siehe Anlage K 56 zur Bereitstellung von Inferenz Workloads von „Hugging Face“), dass dieses Feature überhaupt die Zuweisung von Pods zu Rechen- und Boosterknoten betrifft und inwiefern sich die Zuweisung von Ressourcen zu Projekten und Workloads von der zuvor geschilderten Funktionsweise des Run:ai-Schedulers unterscheidet.

174 Die Klägerinnen tragen insofern unter Berufung auf die Run:ai-Softwaredokumentation (Anlagen K 56) lediglich vor, dass Run:ai jedes Mal, wenn eine variable Bedingung in Bezug auf den zugrundeliegenden Inferenz-Workload (d.h. die Menge der Teilaufgaben) erfüllt sei, einen neuen Replica-Pod (d.h. eine neue Unteraufgabe) erstelle, dem dann Ressourcen – Rechen- und Boosterknoten – zugewiesen würden. Bei den Variablen handele es sich um

Latenz, Durchsatz und Parallelität und somit um Informationen im Sinne von Merkmal 4 des Patentanspruchs.

175 Dieser Vortrag greift zu kurz. Denn die Variablen dienen nicht der Erzeugung einer Verteilung von Unteraufgaben, sondern stellen eine Bedingung für die Erzeugung eines neuen Replikats („replica“) dar. Ein Replikat ist eine Kopie eines Workloads oder eines Pods, die erzeugt wird, wenn ein bestimmter, vom Nutzer gesetzter Grenzwert einer ausgewählten Variable überschritten wird. Die Auswahl einer Variable und die Festsetzung eines Grenzwertes sind überhaupt nur erforderlich, wenn der Anwender optional einen Minimalwert und einen Maximalwert für die Anzahl von Replikaten gesetzt hat, die auf- bzw. abskaliert werden sollen, und diese Werte verschieden sind. Nur dann wird jedes Mal ein Replikat erzeugt, wenn die gewählte Bedingung – der Wert der Variable – eingetreten ist (vgl. Anlage K 56).

176 Mit der Erzeugung eines oder mehrerer Replikate geht jedoch keine (weitere) Verteilung der Pods des „Hugging-Face inference Workloads“ im Sinne von Merkmal 4 des Patentanspruchs einher. Es wird vielmehr lediglich ein Replikat, also eine Kopie eines Pods, erzeugt, und diesem Replikat werden Ressourcen zugewiesen. Von der Vielzahl von Unteraufgaben, die Gegenstand der ersten Verteilung waren, wird durch die Erzeugung eines Replikats im Rahmen des Hugging-Face Inferenz Modells keine Unteraufgabe umverteilt. Sämtliche Pods werden weiterhin von den Knoten bearbeitet, denen sie ursprünglich zugewiesen wurden. Der Wert der Variable, die die Klägerinnen als Informationen im Sinne von Merkmal 4 ansehen, wird demnach nicht verwendet, um eine Umverteilung der Pods gemäß Merkmal 4 zu erzeugen. Eine solche weitere Verteilung findet nicht statt. Stattdessen werden bei Eintritt der Bedingung ein oder mehrere Replikate erzeugt, denen Ressourcen zugewiesen werden müssen. Das Streitpatent verhält sich nicht zur Schaffung weiterer Unteraufgaben. Aber darin besteht jedenfalls nicht die technische Lehre, mit der das Streitpatent das zugrundeliegende technische Problem lösen möchte. Diese Lehre besteht in der weiteren Verteilung der ursprünglichen Unteraufgaben der Rechenaufgabe in Abhängigkeit von Informationen über ihre Bearbeitung in vorhergehenden Iterationen, nicht aber darin, bestehende Unteraufgaben zu replizieren.

c) Iteration

177 Es lässt sich weiterhin nicht feststellen, dass die angegriffene Ausführungsform mit installierter Run:ai-Software geeignet ist, Unteraufgaben in mehreren Iterationen zu berechnen und die Unteraufgaben nach einer ersten Iteration gemäß Merkmal 3 ein weiteres Mal gemäß Merkmal 4 zu verteilen.

178 Die Klägerinnen haben lediglich vorgetragen, jeder Planungszyklus des Run:ai-Scheduler sei eine Iteration im Sinne des Streitpatents. Die Berechnung des Workloads könne unstreitig unterbrochen und währenddessen die für die Berechnung zugewiesenen Ressourcen geändert werden. Diese Unterbrechungen kennzeichneten den Planungszyklus des Run:ai-Schedulers.

179 Dieser Vortrag greift zu kurz. Bei zutreffender Auslegung können die Berechnungen zwischen zwei Planungszyklen nicht ohne weiteres mit einer Iteration im Sinne der Merkmale 3 und 4 gleichgesetzt werden. Der Vortrag der Klägerinnen lässt nicht erkennen, was überhaupt Gegenstand der pods ist und was berechnet wird.

180 Die Beklagten haben zudem bestritten, dass [redacted]. Sie tragen weiter vor, die Unterbrechung der Berechnung der Pods führe nicht dazu, dass die Pods in mehreren Iterationen berechnet würden. Denn die Unterbrechung finde nicht nach einer vollständigen Berechnung/Iteration der Pods statt.

181 Demnach scheint es so, dass die Pods nach einer Unterbrechung weiterberechnet, nicht aber die Rechenanweisung erneut ausgeführt wird. Nach alledem kann jedenfalls nicht zur Überzeugung des Gerichts festgestellt werden, dass die Pods in mehreren Iterationen im Sinne des Streitpatents berechnet und nach einer Iteration umverteilt werden. Dies gilt auch dann, wenn berücksichtigt wird, dass die Pods Aufgaben unterschiedlicher Art betreffen und einzelne Pods im Zeitpunkt der Umverteilung nach einem Planungszyklus beendet sein können und andere Pods umfangreichere Aufgaben betreffen, die noch nicht beendet sind. Nur letztere kommen für eine weitere Verteilung gemäß Merkmal 4 in Betracht. Selbst wenn diese Pods Aufgaben mit iterativem Charakter betreffen, steht die Iteration gleichwohl in keinem Zusammenhang mit der weiteren Verteilung dieses Pods, weil die Verteilung – so die Beklagten in der mündlichen Verhandlung – nach allein zeitlich festgelegten Planungszyklen („scheduling intervals“) unabhängig von jeglicher Iteration und – wie ausgeführt – unabhängig von jeglicher Information in Bezug auf die vorherige Verarbeitung der Unteraufgaben erfolgt.

3. Eignung zur Anwendung des geschützten Verfahrens bei nicht vorinstallierter Software

Auch für den Fall, dass die Run:ai-Software nicht vorinstalliert ist, kann die Eignung der angegriffenen Ausführungsform zur Anwendung des geschützten Verfahrens nicht festgestellt werden. Es ist nicht vorgetragen, dass die Hardware (bloße DGX-System-Produkte ohne Run:ai-Software) – gegebenenfalls ausgestattet mit anderer Software – geeignet ist, das Verfahren nach Anspruch 1 des Streitpatents anzuwenden.

C Widerklage auf Nichtigerklärung

182 Die Widerklage auf Nichtigerklärung bedarf keiner Entscheidung.

183 Die Beklagten haben die Entscheidung über die Widerklage auf Nichtigerklärung davon abhängig gemacht, dass die Verletzungsklage im Übrigen erfolgreich ist. Da Letzteres nicht der Fall ist, bedarf es keiner weiteren Entscheidung mehr über die Widerklage.

184 Der Spruchkörper hält es für prozessual zulässig, dass der Beklagte einer Verletzungsklage eine Widerklage auf Nichtigerklärung erheben und unter die auflösende Bedingung stellen kann, dass die Verletzungsklage keinen Erfolg hat. Tritt diese Bedingung ein, weil das Streitpatent ungeachtet des Rechtsbestands des Streitpatents nicht verletzt ist, ist über die Widerklage auf Nichtigerklärung nicht mehr zu entscheiden.

185 Es kann dahinstehen, ob es sich um eine Klageänderung gemäß Regel 263.3 VerfO, eine Teilklagerücknahme gemäß Regel 265 VerfO oder sogar um eine Form der Erledigung im Sinne von Regel 360 VerfO handelt, wenn der Beklagte erst während des laufenden Verfahrens erklärt, dass die Entscheidung über die Widerklage von dem Ausgang der Verletzungsklage abhängig sein soll (ebenfalls offenlassend: Lokalkammer Mannheim, Entscheidung v. 05.12.2025, UPC_CFI_414/2024 – Centripetal gg. Keysight; für Regel 263.3 VerfO: Lokalkammer München (Panel 1), Entscheidung v. 13.01.2026, UPC_CFI_628/2024 – Emboline gg. Aorticlab). Hat der Beklagte die Widerklage auf Nichtigerklärung vom Ausgang des Verletzungsverfahrens abhängig gemacht, hat der Kläger grundsätzlich kein Interesse daran, dass über die Widerklage auch dann entschieden wird, wenn die Verletzungsklage anderweitig keinen Erfolg hat. Es ist der Beklagte und Widerkläger, der das Rechtsschutzinteresse an einer Widerklage auf Nichtigerklärung formuliert. Dieses liegt jedenfalls dann, wenn der Beklagte wie im vorliegenden Fall die Entscheidung über die Widerklage vom Ausgang des Verletzungsverfahrens abhängig macht, lediglich in der Verteidigung gegen die Verletzungsklage (vgl. Regel 25.1 VerfO). Denn die Widerklage auf Nichtigerklärung ist mit der Verletzungsklage intrinsisch verbunden und regelmäßig die unmittelbare Folge der vom Kläger erhobenen Verletzungsklage (Berufungsgericht, Anordnung v. 20.06.2025, UPC_CoA_393/2025 – Emboline gg. Aorticlab). Hat die Verteidigung gegen die Verletzungsklage anderweitig Erfolg, besteht somit kein Interesse mehr an einer Entscheidung über die Widerklage. Ein solches Interesse außerhalb von Art. 105a EPÜ kann auch der Kläger nicht haben, da er Gefahr läuft, das Patent ungewollt zu verlieren, soweit die Widerklage erfolgreich ist. Aber auch die mögliche Abweisung der Widerklage stellt in der Regel keinen Grund dar, dem Beklagten die bedingte Widerklage zu verweigern. Soweit die Klägerinnen in der mündlichen Verhandlung ihr Interesse bekundet haben, eine für sie positive Entscheidung über die Widerklage anderen potentiellen Gegnern zur Förderung außergerichtlicher Gespräche mitteilen zu wollen, kann dies ein rechtliches Interesse der Klägerinnen nicht begründen. Denn die gerichtliche Entscheidung beschränkt sich in einem solchen Fall auf die Abweisung der Widerklage. Die Rechtsbeständigkeit eines Patents wird nicht positiv festgestellt; auf eine solche Bestätigung des Rechtsbestands ist die Widerklage nicht gerichtet und kann auch vom Kläger und Widerbeklagten nicht verlangt werden. Es kommt hinzu, dass die Abweisung der Widerklage gegenüber Dritten und für andere Gerichte nicht bindend ist. Die bloße Möglichkeit, dass ein potentieller Gegner der Klägerinnen eine Abweisung der Widerklage in außergerichtlichen Gesprächen in seine

Überlegungen einbezieht und es deshalb zu einer außergerichtlichen Einigung kommt, die ein weiteres gerichtliches Verfahren und die damit einhergehenden Kosten vermeidet, mag für die Klägerinnen vorteilhaft erscheinen, begründet aber kein rechtliches Interesse.

186 Ob über die Kosten der Widerklage auf Nichtigerklärung zu entscheiden ist, wenn es zu keiner Entscheidung über die Widerklage kommt, ist bislang nicht höchstrichterlich geklärt (vgl. Lokalkammer Mannheim, Entscheidung v. 05.12.2025, UPC_CFI_414/2024 – Centripetal gg. Keysight; Lokalkammer München (Panel 1), Entscheidung v. 13.01.2026, UPC_CFI_628/2024 – Emboline gg. Aorticlab). Dies bedarf allerdings auch im Streitfall keiner abschließenden Entscheidung, weil die Parteien eine Kostenregelung getroffen haben, die jeder gerichtlichen Kostengrundentscheidung vorgeht. Demnach tragen die Parteien die ihnen durch Verletzungsklage und Widerklage auf Nichtigerklärung entstandenen Kosten jeweils selbst.

ENTSCHEIDUNG

I. Die Verletzungsklage wird abgewiesen.

II. Die Parteien tragen die ihnen durch die Verletzungsklage und die Widerklage auf Nichtigerklärung entstandenen Kosten jeweils selbst. Eine Kostenerstattung findet nicht statt.

III. Der Streitwert der Verletzungsklage wird auf 1.500.000,00 EUR festgesetzt. Der Streitwert der Widerklage auf Nichtigerklärung wird auf 2.250.000,00 EUR festgesetzt.

Dr. D. Voß (Vorsitzender Richter)	
Dr. G. Werner (Rechtlich qualifizierter Richter)	
A. Kupecz (Rechtlich qualifizierter Richter)	

A. Dumont (Technisch qualifizierter Richter)	
Für den Hilfskanzler	

INFORMATIONEN ZUR BERUFUNG

Gegen die vorliegende Entscheidung kann durch jede Partei, die ganz oder teilweise mit ihren Anträgen erfolglos war, binnen zwei Monaten ab Zustellung der Entscheidung beim Berufungsgericht Berufung eingelegt werden (Art. 73 Abs. 1 EPGÜ, Regel 220.1 (a), 224.1 (a) VerfO).

INFORMATIONEN ZUR VOLLSTRECKUNG

Die Entscheidung hat keinen vollstreckbaren Inhalt.

Die Entscheidung wurde am 11. März 2026 in öffentlicher Sitzung verkündet.

Vermerk

Bei diesem Dokument handelt es sich um die um vertrauliche Informationen bereinigte, redigierte Version der Entscheidung. Sie ist ohne die Unterschriften der beteiligten Richter und des Vertreters des Hilfskanzlers gültig.

Daniel Voß

Digital unterschrieben von
Daniel Voß
Datum: 2026.03.11 11:40:51
+01'00'