Quartl

2/2023
106. Ausgabe

Inhalt

Das Quartl erhalten Sie online unter `https://www.cs.cit.tum.de/sccs/weiterfuehrende-informationen/quartl/`

KONWIHR

Elitenetzwerk Bayern

Das Quartl ist das offizielle Mitteilungsblatt des *Kompetenznetzwerks für Technisch-Wissenschaftliches Hoch- und Höchstleistungsrechnen in Bayern* (KONWIHR) und der *Bavarian Graduate School of Computational Engineering* (BGCE)

Editorial

Dieses Mal habe ich wieder einen Lesetipp der Extraklasse für Sie: www.seelenglueck.at. Wie ich nun wieder an dieses Highlight komme? Nun, gemeinsam mit einem Mitarbeiter auf dem Weg zu einem Workshop in Österreich, fiel mir beim Halt an der Raststätte Holzkirchen (zum Erwerb des obligatorischen Wegelagerer-Pickerls) ein Auto mit besagter URL auf. Nach fast 30 Jahren Quartl sind meine Sensoren offensichtlich entsprechend empfindlich eingestellt. Auf der Weiterfahrt zückt der Mitarbeiter das Handy und schaut sich die Webseite an. Bereits nach den ersten vorgelesenen Sätzen war klar: Die Spürnase hatte Recht, das muss ins Quartl!

Hinter dem Seelenglück made in Austria steht die Firma einer gewissen Bettina G., nach eigener Auskunft Human- und Tierenergetikerin von Beruf, was auch immer das sein soll. Ein Blick auf ihrem Werdegang zeigt, dass die Dame eine in Deutschland ausgebildete Heilpraktikerin ist. Nun ist es bekanntlich so, dass sowohl die Ausbildung hierzu als auch die Ausübung dieses Berufs in Österreich gesetzlich verboten sind, aus welchen Gründen auch immer. Das erklärt vielleicht das mit der Tierenergetikerin etwas .....

Doch schauen wir in der Menüleiste der Webseite, was das Portfolio des österreichischen Seelenglücks so zu bieten hat: Da finden sich klangvolle Begriffe wie „Mediales Coaching“, „Geistiges Heilen“ und „Jenseitskontakte“, im Angebot jeweils einstündige Sitzungen à 70 Euro. Oder dann „Tierkommunikationünd „Lichtkommunikation“. Schauen wir uns letztere mal etwas genauer an. Da steht unter anderem: *„Mithilfe der heiligen Geometrie wird die göttliche Ordnung in dir wieder neu zum Leben erweckt und es können sich neue Wege und Türen für persönliches Wachstum und Balance öffnen.*

Oh heilige Geometrie! Ohne jetzt zu sehr ins Detail gehen zu wollen, würde mich mal die Performance von Bettina G. in Schule oder Ausbildung in Geometrie interessieren. Das Ganze gibt es übrigens in der Form eines „persönlichen Gebets“ für 44 Euro oder, falls man etwas mehr investieren möchte, in Form eines „Manifestations-, Heilgebets“ für 66 Euro.

Doch es gibt noch mehr – man kann sich die Energie auch aufsprühen lassen. Da wäre zum Beispiel das „Seelenauraspray“, von dem es heißt: *Für dein persönliches Auraspray verwende ich hochwertige ätherische Öle, Heilsteine, Essenzen der Lichtkommunikation und speziell für dich übermittelte Heilgebete aus der lichtvollen geistigen Welt.* What the f*** sind Essenzen der Lichtkommunikation? Ehrlich – einen kurzen Moment fragte ich mich, ob es die versteckte Kamera auch in einer Online-Version gibt ... Ihr Seelenauraspray bekommen Sie übrigens gegen Elemente der Geschäftskommunikation in Höhe von 55 Euro. Dass es auch noch ein Themenauraspray gibt (je nach Packungsgröße für zwischen 11 und 33 Euro), sei hier nur am Rande erwähnt.

Über sich schreibt Bettina G. übrigens: *Ich lache gerne und viel und lasse die Sonne im Herzen anderer scheinen.* Gelacht haben wir noch im Auto in der Tat viel. Insofern stimmt das mit der Sonne irgendwie – auch wenn sich Bettina G. wahrscheinlich nicht als Komikerin sieht. Aber im Grunde ist das alles nicht so lustig. Denn neben gelangweilten Reichen werden sich beim Seelenglück auch Verzweifelte einfinden, die auch willens sind, ihr Heil in der Lichtkommunikation oder im Seelenauraspray zu suchen – weil sie nach jedem Rettungsanker greifen. Eigentlich eher schamlos, da Sprays zu verkaufen.

Einen hab ich noch, der mir dankenswerterweise wieder von einem treuen Quartl-Leser zugespielt wurde: Wussten Sie schon von Erich Honeckers umfassender Expertise in Sachen Kultur? Falls nicht, einfach mal googeln – da findet sich bei Youtube ein absolutes Kleinod: *Über eine lange Zeit hat DT64 (Jugendrundfunksender in der ehemaligen DDR; Anm. der Redaktion) in seinem Musikprogramm einseitig die Beatmusik propagiert. Hinzu kam,*

dass im Zentralrat der Freien Deutschen Jugend eine fehlerhafte Bewertung der Beatmusik war. Sie wurde als musikalischer Ausdruck des Zeitalters der technischen Revolution entdeckt. Dabei wurde übersehen, dass der Gegner diese Art von Musik ausnützt, um durch die Übersteigerung der Beatrhythmen Jugendliche zu Exzessen aufzuputschen. Der schädliche Einfluss solcher Musik auf das Denken und Handeln von Jugendlichen wurde grob unterschätzt. Niemand in unserem Staate hat etwas gegen eine gepflegte Beatmusik (ebenso, wie niemand die Absicht hatte, eine Mauer zu bauen; weitere Anmerkung der Redaktion), sie kann jedoch nicht als die alleinige und hauptsächliche Form der Tanzmusik betrachtet werden.

Mei, Erich, hättest du halt die Fortbildungsangebote von Udo & Co. (*Hey Honey, ich sing für wenig Money ...*) angenommen, im Sinne eines Life-long Learning. Aber es war bekanntlich ja nicht das einzige Mal, wo du nicht so ganz richtig lagst ...

Doch damit genug für heute. Die gesamte Quartl-Redaktion wünscht Ihnen einen schönen Sommer und eine erholsame Ferienzeit – mit oder ohne Beatmusik, aber definitiv mit viel Seelenglück, auch ohne Bettina G. Zunächst aber wünschen wir Ihnen natürlich viel Vergnügen mit der neusten Ausgabe Ihres Quartls!

Hans-Joachim Bungartz.

Iterationsschleife

N=47
12. Juni 2023

Istrien[a] und die Stadt Triest[b] waren im 20. Jahrhundert lange umstritten. Grenzziehungen haben lange Zeit das Verhältnis der Stadt zu ihrer Umgebung und insbesondere zu Istrien geprägt. Im Jahr 1900 bildeten Triest, Istrien und die Gefürstete Grafschaft Görz und Gradisca[c] die österreichischen Küstenlande. Diese waren für die Habsburg Monarchie von herausragender Bedeutung.

Triest war ein wichtiger Hafen. 1829 führte dort der österreichische Erfinder Josef Ressel[d] die erste Fahrt eines Schiffes mit einer Schiffsschraube durch. 1831 wurde in Triest die Assicurazioni Generali[e] gegründet. 1833 wurde der Österreichische Lloyd gegründet. 1857 wurde Triest über die Südbahn mit Österreich verbunden. So blühte die Stadt Ende des 19ten Jahrhunderts auf. Die Stadt Triest selbst war überwiegend italienisch geprägt mit einer großen deutschsprachigen Minderheit. Das Umland war friaulisch, slowenisch und kroatisch geprägt. Die Industrialisierung des ausgehenden 19ten Jahrhunderts konnte also auf ein großes Arbeitskräftereservoir aus dem Umland zurückgreifen, was aber auch zu nationalistischen Debatten in der Stadt führte.

---

[a] Istrien ist eine große Halbinsel in der Adria. Der Name geht wohl auf das Volk der Histrier zurück, deren Zugehörigkeit unter Historikern umstritten ist. Ironischerweise ist unklar ob die Histrier eher zu den Venetern (heute Italien) oder zu den Illyrern (Jugoslawien bzw. Slowenien/Kroatien) gehörten.

[b] Triest ist eine norditalienische Hafenstadt mit ca. 200.000 Einwohnern. Der Ort ist seit dem 2. Jahrtausend vor Christus bewohnt. Seinen Namen bekam die Stadt vermutlich von den Venetern die die Stadt vor rund 3000 Jahren Tergeste nannten. Julius Cäsar erwähnt die Übernahme der Stadt in das römische Reich in seinen Kommentaren zum gallischen Krieg.

[c] Die Grafschaft war seit 1500 im Besitz der Habsburger und bis 1918 Kronland der Habsburg Monarchie.

[d] Josef Ludwig Franz Ressel (1793 – 1857) war ein österreichischer Erfinder. Er erfand unter anderem die Schiffsschraube, war aber nicht erfolgreich darin, diese Erfindung auch unter seinem Namen zu vermarkten. Er blieb in Österreich und starb verbittert. Die Schiffsschraube machte in England Karriere.

[e] Die Assicurazioni Generali ist noch heute als Generali bekannt. Sie hat aktuell rund 75.000 Mitarbeiter:innen und einen Jahresumsatz von rund 100 Mrd. Euro. Literaturverliebten ist die Assicurazioni Generali als erster Arbeitgeber von Franz Kafka (1883 – 1924) bekannt, für die Kafka von Oktober 1907 bis Juli 1908 tätig war. In dieser Zeit entstanden die „Hochzeitsvorbereitungen auf dem Lande“.

Gleichzeitig war Triest ein Schmelztiegel verschiedener Kulturen, der auch internationale Anziehungskraft besaß. [ab]

1919 kamen aufgrund der Geheimverträge über den Eintritt Italiens in den ersten Weltkrieg[c] Triest und Istrien zu Italien. Die Stadt Rijeka/Fiume auf der Ostseite Istriens wurde zum Freistaat erklärt.[d] Nach dem zweiten Weltkrieg verloren die Italiener Istrien aber beinahe völlig. Insgesamt wurden von Jugoslawien etwa 300.000 Italiener ausgewiesen und mussten nach Italien zurückkehren. Triest wurde mit seinem Umland als freies Territorium Triest 1947 unter die Oberhoheit der UN gestellt.[e] 1954 wurde das Territorium aufgeteilt.[f] Triest verblieb bei Italien. Sein Umland kam an Jugoslawien. Triest verlor damit seine Verbindung zum istrischen Hinterland und seine wirtschaftliche Bedeutung. Etwa 20.000 Triestiner wanderten zwischen 1954 und 1961 aus. 1975 wurde die bis heute gültige Grenze zwischen Slowenien und Italien im Vertrag von Osimo festgelegt. Mit dem Ende Jugoslawiens und dem Beitritt Sloweniens (2004) und Kroatiens (2013) zur Europäischen Union sind Istrien und Triest heute wieder eng miteinander verbunden. Einen wirtschaftlichen Aufschwung hat die Stadt dennoch nicht erlebt. Aber die Stadt und ihr Hafen haben das Interesse Chinas geweckt, das Triest gerne in sein Projekt der neuen Seidenstraße einbinden will. Touristisch blüht die Region dagegen gerade auf. In den westlichen Küstengegenden Istriens und in Triest hört und spricht man wieder Kroatisch, Slowenisch, Italienisch, Deutsch und Englisch. Istrien und Triest sind wieder international wie 1900, und sie sind wieder von Nationalismen bedroht - wie 1900.

M. Resch

---

[a] Siehe dazu: Claudio Magris, Angelo Ara „Triest: eine literarische Hauptstadt in Mitteleuropa“, dtv, 2005 sowie Scipio Slataper „Mein Karst und andere Schriften“, Promedia, 1988

[b] Unter anderem lebte der irische Schriftsteller James Joyce (1882 – 1941) von 1905 – 1915 in Triest, nachdem er zuvor ein Jahr im Habsburgischen istrischen Kriegshafen Pula/Pola verbracht hatte. In Triest traf Joyce den italienischen Schriftsteller Ettore Schmitz (1861 – 1928 besser bekannt als Italo Svevo), der zum wesentlichen Vorbild für Leon Blum, den Helden des joyce‘schen Hauptwerks Ulysses, wurde.

[c] Londoner Vertrag vom 26. April 1915 geschlossen zwischen Italien einerseits und Großbritannien, Frankreich und Russland andererseits.

[d] Der italienische Schriftsteller Gabriele d’Annunzio (1863 – 1938) war damit nicht einverstanden und besetzte Fiume/Rijeka am 12. September 1919 mit 2500 italienischen Freischärlern. Nach längeren Streitigkeiten einigten sich das Königreich der Serben, Kroaten und Slowenen (später Königreich Jugoslawien) und Italien am 12. November 1920 im Grenzvertrag von Rapallo darauf, Fiume als Freistaat zu garantieren. Schon 1924 wurde Fiume wieder italienisch (Vertrag von Rom 27.1.1924).

[e] Vertrag von London geschlossen 1947 zwischen Italien und den Alliierten

[f] Londoner Abkommen geschlossen 1954 zwischen Italien und Jugoslawien

Zeitintegration: Hochskalierbare rationale Approximation von exponentiellen Integratoren

Forschung im Überlappungsbereich von Höchstleistungsrechnern und Lösern für Partielle Differentialgleichungen (PD) behandelt leider nur zu häufig die räumlichen Aspekte der Differentialgleichungen. Hierbei scheint es mir so, als ob die Zeitdimension an sich zwar nicht gänzlich, aber dennoch sehr stark vernachlässigt wurde. An dieser Stelle möchte ich die Gelegenheit nutzen, um die Leser auf neue Forschungsgebiete der letzten 10 Jahre aufmerksam zu machen und in diesem Artikel auf "REXI".

Ich kann mich nicht erinnern, dass mich etwas mehr fasziniert hat als die Möglichkeit, eine lineare autonome partielle Differentialgleichung in der Zeit ohne jegliche CFL Einschränkung integrieren zu können. Da lernt man zuerst in seiner Ausbildung, dass es steife Probleme gibt und dass die CFL Nummer ein immer limitierender Faktor ist. Und dann lernt man plötzlich eine Methode kennen, mit der sich diese Eigenschaften unter bestimmten Umständen aushebeln lassen.

Gegeben sei eine PD in der semi-diskreten Form

$$\frac{\partial}{\partial t} U(t) = LU(t)$$

wobei U ein Zustandsvektor zur Zeit t ist, $\frac{\partial}{\partial t} U$ die Zeitableitung von U und ein linearer diskreter Operator L (eine Matrix).

Traditionelle Zeitintegrationsverfahren versuchen nun Schritt-für-Schritt die Lösung zu berechnen, indem man von einem Zustand $U(t_0)$ zu einem gewissen Zeitpunkt t_0 ausgeht und anschliessend mit einer Matrix-basierten Padè Approximation die Lösung über einen gewissen Zeitschritt hinweg approximiert. Hierzu gibt es nahezu beliebig viele Methoden (Runge-Kutta Butcher table, Leap-frog, spectral deferred corrections, ..., etc.) und alle

basieren auf der Idee, ein Zeitintervall basierend auf (iterativ) berechneten Zwischenwerten zu approximieren.

Ein disruptiver Ansatz stellt die **rationale Approximation** von exponentiellen Integratoren dar. Hier sei gleich zu Anfang gesagt, dass diese Namensgebung der einfachen Tatsache entspringt, dass man den Zustandsvektor U zu einem beliebigen Zeitpunkt durch

$$U(t+\Delta t) = \sum_i \frac{\beta_i}{\Delta t L - I\alpha_i} U(t) = \sum_i \beta_i \left(\Delta t L - I\alpha_i\right)^{-1} U(t) \quad (1)$$

berechnen kann, wobei α_i und β_i komplexe Zahlen sind. Warum ist aber diese Form möglich - und wieso verwendet's niemand?! Zuerst eine kleine Herleitung: Mittels "Diagonalisierung" gelangen wir auf eine Form

$$\tilde{u}(t) = e^{\Delta t \lambda} \tilde{u}(t_0)$$

wobei λ einen der Eigenwerte der L Matrix darstellt und $\tilde{u}$ das dazugehörige Element des Zustandsvektors im Eigenraum (aber eine Eigenwertszerlegung werden wir am Ende nicht benötigen!).

Offensichtlich leidet diese Form weder an der CFL noch an anderen restriktiven Problemen von impliziten Verfahren, da die exponentielle Funktion direkt ausgewertet werden kann. Um diesen Ansatz nun auf die PD anwenden zu können, kann man eine bestimmte Approximation (ich verweise hier indirekt auf Jahrzehnte von Literatur) für die exponentielle Funktion verwenden, welche es erlaubt, die Lösung wieder als $U(t) = ...$ zu schreiben. Dies ist möglich mittels rationaler Terme der Form

$$e^{\Delta t \lambda} = \sum_i \frac{\beta_i}{(\lambda - \alpha_i)}$$

oder anders ausgedrückt: Einfach eine Linearkombination von rationalen Basisfunktionen - nur eben nicht im Raum wie wir es kennen sondern einfach mal in der Zeit. Der Leser kann sich mittels einfacher Algebra nun selbst

davon überzeugen, dass man wieder zurück zu der Form von Gleichung 1 gelangen kann. Nur ist dabei zu beachten, dass die rationale Approximation nun alle Eigenwerte von L approximieren sollte (Kommentare zu Filteransaetzen ueberspringe ich hier mal...).

So, wozu denn nun dieser ganze Kram? Der Grund ist sehr einfach: Man kann nun jeden Term der Summe **parallel rechnen und genau hier kommen Höchstleistungsrechner in's Spiel**. Zuerst wird der momentane Zustand $U(t)$ auf die Rechenressourcen verteilt, anschliessend das Gleichungssystem

$$\beta_i \left(\Delta t L - I\alpha_i\right)^{-1} U(t)$$

gelöst (wobei ich hierbei darauf hinweisen sollte, dass keine out-of-the-Box Löser wie Bi-CG, GMRES oder Multigrid funktionieren weshalb hier noch mehr oder weniger Forschung nötig ist) und mittels einer elementweisen Reduzierung der neue Zustand $U(t + \Delta t)$ berechnet.

Fertig ist unser CFL-freier Löser für lineare autonome PDs. Ein grösserer Zeitschritt bedeutet hierbei lediglich mehr Rechenressourcen zu nutzen, statt wie mittels traditioneller Methoden mehr Zeitschritte zu berechnen. Auch die Genauigkeit kann durch die Verwendung von mehr Rechenressourcen einfach erhöht werden. Alles in allem also eine wunderbare Methode, oder? Wenn da nicht die nicht-Linearitäten wären... Aber darüber gibt's vielleicht mal einen anderen Artikel.

Martin Schreiber

Die Sprache des Lebens verstehen und sprechen lernen

Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

Das CS-2-System von Cerebras Systems am Leibniz-Rechenzentrum unterstützt Bioinformatiker:innen dabei, die Codes von Proteinen zu entschlüsseln. Damit können neue Heilverfahren entstehen oder drängende Umweltfragen gelöst werden.

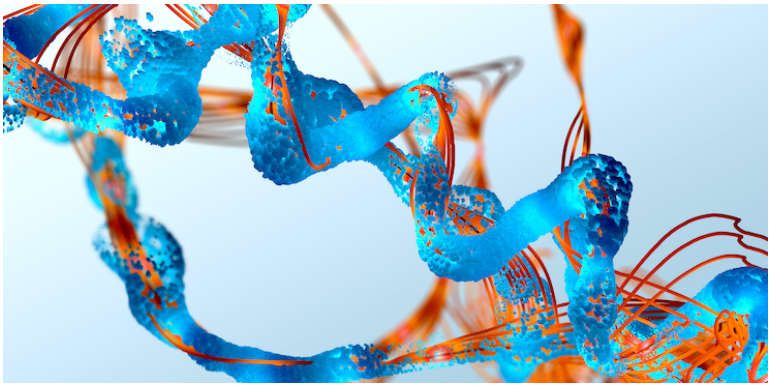


Abbildung 1: Proteine: Ihr Aufbau ähnelt der einer Sprache. Mit Sprachmodellen kann ihr Code entschlüsselt werden.

Proteine sind der Grundstoff des Lebens. Sie bestimmen Form, Aufbau und Funktionen von Zellen, Gewebe, Organen sowie den Stoffwechsel und das Wachstum, und zwar von Mensch, Tier, Pflanze. Ein Großteil von ihnen setzt sich aus etwa 20 Aminosäuren zusammen, so ein Ergebnis von rechnergestützten Analysen der letzten Jahrzehnte. Zwar sammeln sich inzwischen in vielen Datenbanken Milliarden von Aminosäure-Kombinationen und Variationen von Proteinsequenzen, doch wie diese Eiweißstoffe für Wachstum sorgen oder Zellfunktionen beeinflussen, ist noch immer weitestgehend ein Rätsel.

Und: Die Suche in den prall gefüllten Datenbanken dauert viel zu lange. Doch Künstliche Intelligenz (KI) und Mustererkennung, neuerdings vor allem Modelle zur Verarbeitung menschlicher Sprache helfen, den Code des Lebens zu knacken. Sie beschleunigen zudem die Suche: „Sprachmodelle lernen Muster und Ähnlichkeiten in den Sequenzen direkt aus Proteindatenbanken", erklärt Dr. Michael Heinzinger, Bioinformatiker der Technischen Universität München (TUM) und Mitarbeiter des Rost-Labs am Lehrstuhl Bioinformatik und numerische Biologie von Prof. Burkhard Rost. „Die traditionelle, statistische Mustererkennung dauert in der Regel lange und funktioniert nicht mit allen Proteinen gleich gut. Sprachmodelle verkürzen den Such- und Analyseprozess und geben uns neue Werkzeuge an die Hand, um das Verständnis von Proteinen zu verbessern."

Sprachmodelle entschlüsseln den Protein-Code

Die Suche nach bestimmten Proteinstrukturen anhand von wiederkehrenden Mustern kann Tage, wenn nicht sogar Wochen dauern. Vor etwa vier Jahren entdeckte das Team um Prof. Burkhard Rost die Analogie von menschlicher Sprache und Proteinen: Die 20 wichtigsten Aminosäuren funktionieren wie Buchstaben, fügen sich quasi zu Wörtern und Sätzen, besser zu Proteinen und -Sequenzen mit eigenen Funktionen. Diese müssten, so eine Annahme am Rost-Lab, folglich mit smarten Programmen zur Verarbeitung natürlicher Sprache zu entschlüsseln sein. Statt wie gewohnt mit Artikeln von Wikipedia wurde daher ELMo, das Embeddings from Language Model, mit Proteinsequenzen gefüttert und trainiert. Das Ergebnis war SeqVec (Sequence-to-Vector), das erste smarte Modell zur Verarbeitung von Protein-Codes, das Datenbanken nicht mehr nach Mustern durchsucht, sondern direkt den Protein-Code aufnimmt und diesen eigenständig auf neue Proteine überträgt. Nach dessen Vorbild trainierten Forschende weltweit mit Large Language Models (LLM) und entwickelten bessere, effizientere Trainingsmodelle für Proteine mit neuen Funktionalitäten. Neue Technologien für die Verfahren der KI befeuerten diese Entwicklung.

Insbesondere die sogenannten Transformer – mit ihrer Hilfe wandelt der Computer Buchstaben- und Zeichenfolgen in mathematische Vektoren um

– erweiterten die Möglichkeit des Trainings mit NLP-Modellen. Für die Protein-Analyse veröffentlichte der TUM-Lehrstuhl Bioinformatik verschiedene, auf Proteinsequenzen trainierte Transformer, namens ProtTrans. Diese ersten Proteincode-Modelle entstanden zunächst am eigenen Computercluster des Lehrstuhls sowie an den AI-Systemen des Leibniz-Rechenzentrums (LRZ), also auf parallel geschalteten Graphics Processing Units (GPU), speziell fürs maschinelle Lernen optimierte Prozessoren. Seit vergangenem Jahr steht den Wissenschaftler:innen am LRZ noch ein Cerebras CS-2 System zur Verfügung. „Ein völlig anderes KI-System", sagt Heinzinger. „Mit seinem großen Chip und der hohen Speicherkapazität vereinfacht es viele Arbeitsschritte des verteilten Trainings. Wir müssen uns beispielsweise nicht mehr um die Kommunikation zwischen Prozessoren und Knoten kümmern." Modelle könnten schneller optimiert werden, nach ersten Erfahrungen von Heinzinger und seinen Kolleg:innen wird auch das Training mit großen Datenmengen beschleunigt. Allerdings fielen grundlegende Veränderungen oder Neuerungen an einem Modell vergleichsweise schwer. „Das ist nur eine Frage von Zeit, das CS-2-System erfordert für den Einsatz in der Forschung weitere Komponenten, und mit diesen kommt ein überarbeiteter Software Stack", sagt Nicolay Hammer, promovierter Astrophysiker und Leiter des Teams Big Data & AI am LRZ.

Ging es bislang vor allem ums Kennenlernen von Protein-Sequenzen, den Wörtern des Lebens-Codes, sollen die neuen, besseren und spezialisierten Sprachmodelle nun auch die Syntax und Grammatik des Protein-Codes dechiffrieren und aufzeigen, wie und wofür Eiweißstoffe dreidimensionale Strukturen bilden, sich falten und was das bewirkt. „Wir können mit größeren Sprachmodellen nicht nur die Proteine besser verstehen, sondern diese auch zielgerichtet manipulieren oder neu schreiben, um damit den Herausforderungen des 21. Jahrhunderts zu begegnen", ergänzt Heinzinger Möglichkeiten. „Durch ihr breites Funktionsspektrum sind Proteine in vielen pharmakologischen und biotechnologischen Prozessen unentbehrlich." Mit ihnen werden zum Beispiel Medikamente hergestellt, neuerdings auch Biokraftstoffe oder Materialien, die Plastik abbauen oder Kohlenstoffe binden. Und wer den Code der Proteine beherrscht, kann damit unbegrenzt neue Moleküle oder

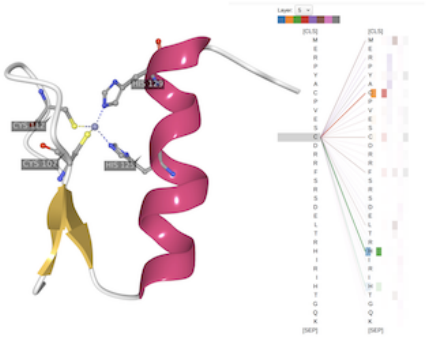


Abbildung 2: Die Grafik verdeutlicht die Funktionsweise von Sprachmodellen, die Kennungen für Aminosäuren wie Buchstaben zu Sequenzen oder Wörtern fügen.

Stoffe schaffen.

Rechenkraft verkürzt das Training

Sprach- und Proteinstrukturen ähneln sich, doch es gibt natürlich auch Unterschiede. Um Computern und smarter Sprachverarbeitung die Eigenheiten von Proteinen zu vermitteln, konfrontierten die Spezialist:innen des Rost-Labs erst ELMo und weitere Transformer-basierte Sprachmodelle mit Datensets, die zwischenzeitlich bis zu 2,3 Milliarden und mehr Proteinsequenzen enthalten. Normalerweise registrieren Sprachmodelle Buchstaben- und Wortkombinationen, in diesem Fall aber, wann sich welche Aminosäuren aneinanderreihen. Ähnlich einem Lückentext in der Schule füllen die Programme danach künstlich erzeugte Leerstellen und beweisen damit, dass sie den Aufbau der Proteine nachvollziehen können. Sucht das menschliche Hirn mit Intuition und nach Sinn fehlenden Wörtern, sortiert die Maschine Lösungen nach statistischer Verteilung. Sie nennt verschiedene Variationen für die Lücke nach den Wahrscheinlichkeiten, mit der diese in Proteinen auftreten. „Je besser Computer eine Proteinsequenz lesen können, umso besser können sie deren Strukturen und Funktionen verstehen“, so Heinzinger. Schrittweise werden danach Parameter, die Kriterien für Auswahl und Analyse, verändert oder

hinzugefügt.
Mehrere Dutzend Trainingsläufe sind die Regel. Mit jedem Parameter wachsen zudem die Modelle und neuronalen Netzwerke: Enthielt SeqVec 93 Millionen Neuronen, enthält das jüngste Modell aus dem Rost-Lab ProtT5 schon drei Milliarden: „Ein Sprachmodell zu trainieren, ist teuer, es dauert mit jedem Durchlauf länger, weil die Datensätze, aber auch die Zahl der Parameter wachsen. Das braucht Rechenkraft und Energie“, sagt Bioinformatiker Heinzinger. „Aber danach bekommen Anwender:innen aus Biotechnologie, Pharmakologie oder Medizin im besten Fall ein brauchbares Modell für ihre eigenen Analysen an die Hand.“

Je nach verfügbarer Rechenleistung und Größe der Modelle dauern Trainingsläufe Stunden oder mehrere Wochen. Dabei werden Daten immer wieder durch Prozessoren und Speichereinheiten geschleust, neu kombiniert, ausgewertet, überprüft, abgespeichert, weiterverarbeitet. Stehen im LRZ AI-Cluster 68 parallel geschaltete GPU mit dynamischem Arbeitsspeicher (DRAM) von jeweils zwischen 16 und 80 Gigabyte zur Verfügung, bietet das Cerebras-System einen einzigen Chip, auf dem sich insgesamt 850.000 Arbeitseinheiten rund 40 Gigabyte Speicher teilen. Dieser ist auch noch mit HPE Superdome Flex Servern verbunden, die weitere 10 Terabyte Arbeits- (RAM) sowie 100 Terabyte Datenspeicher mitbringen. So kommt eine Rechenleistung zusammen, für die sonst Dutzende von GPU erforderlich wären. Daten können auf diesem Supercomputer blitzschnell zwischen Arbeits- und Speichereinheiten fließen. „Ein Training mit ProtT5 wäre auf den anderen LRZ AI-Systems nicht mehr möglich, es würde Jahre dauern“, sagt Heinzinger. „Im Gegensatz zur Mustererkennung, die für jedes neue Protein eine Datenbank immer wieder durchsuchen müsste, trainieren Sprachmodelle direkt auf den Rohdaten einer Datenbank. Die so aufgenommenen Proteincodes können danach direkt auf neue, andere Proteine angewandt werden.“
Zurzeit versucht der Bioinformatiker, unterschiedliche, etablierte LLM auf dem Cerebras CS-2-System zu implementieren und deren Tauglichkeit für Fragen aus Biologie und anderen Lebenswissenschaften zu überprüfen. Es geht dabei zunächst aber weniger darum, Proteine zu analysieren als die neue Technologie in den Griff zu bekommen und das System auf eigene Ansprüche

anzupassen.

„Sprachmodelle eröffnen neue Möglichkeiten, etwa die Generierung von Texten oder Proteinsequenzen“, so Heinzinger. Spezialist:innen können nicht nur neue Stoffe, Materialien oder Medikamente mit den dechiffrierten Proteinsequenzen aufbauen, einmal austrainiert bieten Modelle wie ProtT5 außerdem Grundlagen für die Entwicklung von Software: etwa zur Analyse von Organismen und Gewebe, zum Erkennen und Behandeln von Mutationen oder Krankheiten, auch zur Entwicklung und Gestaltung neuer organischer Stoffe und Materialien. Auch dafür lohnt sich der hohe Aufwand des Trainings.

Susanne Vieser

Apply now!

The competence network for scientific high-performance computing in Bavaria accepts applications for “normal” (up to 12 months, 50 000 €) and “small” (up to 3 months, 10 000 €) projects, as well as “basis” projects to establish contact points.

The main objective of KONWIHR is to provide technical support for the use of high-performance computers and to expand their deployment potential through research and development projects. Close cooperation between disciplines, users, and participating computer centers as well as efficient transfer and fast application of the results are important.
Find already funded projects at

`https://www.konwihr.de/konwihr-projects/`

Read more on konwihr.de and join the konwihr-announcements mailing list.

Contact KONWIHR

For any KONWIHR inquiries, you only need one address:

info@konwihr.de

Your email will be read carefully and answered by Katrin Nusser or Gerasimos Chourdakis, KONWIHR’s current contact people in the Bavarian North and South. Together with Prof. Gerhard Wellein and Prof. Hans-Joachim Bungartz (who you can also reach using the same address), we collect and process your proposals two times per year (1st of March and 1st of September). Learn more about how you can apply for funding at:

`https://www.konwihr.de/how-to-apply/`

Katrin Nusser, Gerasimos Chourdakis

New projects from spring 2023

The competence network for scientific high-performance computing in Bavaria welcomes the new projects that succeeded in the application round of spring 2023. In every round, we accept proposals for "normal" (up to 12 months) and "small" (up to 3 months) projects, as well as "basis" projects to establish contact points. In this round, the following projects were funded:

- *Unbinned analysis framework for Gammapy*
  by Prof. Stefan Funk and Tim Unbehaun,
  Erlangen Centre for Astroparticle Physics (ECAP), FAU
- *Performance-Optimierung und Parallelisierung eines Codes zur Lösung von partiellen Differentialgleichungen auf dünnen Gittern*
  by Prof. Christoph Pflaum and Riccarda Scherner-Grießhammer,
  Informatik 10, FAU

You can find more details about these projects at

`https://www.konwihr.de/konwihr-projects/`

We would like to invite you to our online workshop on October 11, 15:00-17:00, in which new projects will present their goals and challenges. For more details, see `https://www.konwihr.de/`.

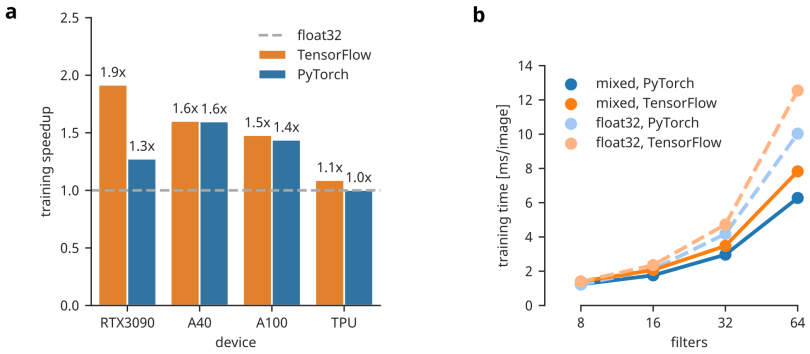
Katrin Nusser, Gerasimos Chourdakis

Projektabschluss: HPC Mixed Precision Quantization of Encoder-Decoder Deep Neural Networks

KI-Modelle werden aktuell immer komplexer und leistungsfähiger. Deshalb wird auch ihr Training immer rechenintensiver. Üblicherweise werden hierfür GPUs oder spezielle Hardware genutzt. Neuronale Netze werden oft viele Male auf großen Datenmengen trainiert, was zu einem hohen Energieverbrauch und CO2-Ausstoß führt.

Bei Berechnungen setzen KI-Frameworks wie TensorFlow und PyTorch standardmäßig auf Gleitkommazahlen mit 32-Bit Genauigkeit. Wenn ein Netz stattdessen mit Mixed Precision trainiert wird, also mit einer dynamischen Mischung von Gleitkommazahlen in 32-Bit und 16-Bit Genauigkeit, kann das Training beschleunigt und dadurch der Energieverbrauch gesenkt werden. Das liegt daran, dass moderne GPUs mit Tensor Cores ausgestattet sind, spezielle Recheneinheiten für Matrixmultiplikationen in 16-Bit. Bestimmte Parameter werden allerdings in 32-Bit gespeichert, um die Genauigkeit nicht zu beeinträchtigen.

In unserem KONWIHR-Projekt am Department of Artificial Intelligence in Biomedical Engineering (AIBE) an der FAU Erlangen-Nürnberg haben wir den Einsatz von Mixed Precision im Anwendungsfall der medizinischen Bildsegmentierung untersucht. Dazu haben wir ein neuronales Netz auf dem öffentlich zugänglichen Datensatz BAGLS (Benchmark for Automatic Glottis Segmentation) trainiert, welcher Bilddaten enthält, die bei der Highspeed-Videoendoskopie des Kehlkopfes aufgenommen wurden [1]. Unsere Experimente haben wir mit einer Netzarchitektur durchgeführt, die auf dem U-Net [2] beruht. Wir haben die Auswirkung von Mixed Precision auf die Geschwindigkeit des Trainings und der Inferenz in unterschiedlichen Szenarien und sowohl auf HPC-GPUs als auch auf TPU-Geräten verglichen.



Vergleich der Trainingsgeschwindigkeit in verschiedenen Szenarien. a) Beschleunigung durch die Mixed-Precision Methode auf verschiedenen GPUs und einer Cloud-TPU, verglichen mit 32-Bit-Training. b) Die Verarbeitungszeit pro Bild und der Beschleunigungseffekt stieg mit der Größe des Netzwerks.

Unsere Experimente haben bestätigt, dass Mixed Precision das Training auf allen untersuchten GPUs beschleunigen konnte und die Segmentierungsqualität nicht beeinträchtigt hat. Die Beschleunigung war abhängig von der GPU, wie in Abbildung a) zu sehen ist. Beispielsweise konnte auf einer Nvidia RTX3090 bis zu 1,9x schneller trainiert werden. Auf der HPC-GPUs A40 und A100 war das Training etwa 1,4x bis 1,6x schneller.
Abbildung b) zeigt, dass der positive Effekt auf die Geschwindigkeit mit zunehmender Modell-Komplexität (repräsentiert durch die Anzahl initialer Filter der Faltungsschichten) größer wurde. Auch das Framework und die Batchgröße hatten einen Einfluss auf die Beschleunigung. Wie in Abbildung a) zu sehen ist, hatte die Aktivierung von Mixed Precision kaum einen Einfluss auf die Geschwindigkeit bei einer Cloud-TPU von Google, da diese intern bereits dynamisch den Datentyp „bfloat16“ nutzt. Insgesamt war die TPU jedoch deutlich schneller als die GPUs.

Wir haben herausgefunden, dass auch die Anwendung eines bereits trainierten Netzes, die Inferenz, von dieser Methode profitiert. Hier war der Effekt aber etwas kleiner. Für die Inferenz von kleinen KI-Modellen ist die Edge-TPU eine energieeffiziente, platzsparende und günstige Alternative zu GPUs. Für die Ausführung auf der Edge-TPU mussten die trainierten Netze zunächst zu 8-Bit Integer quantisiert werden, was zu einer geringen Reduzierung der Segmentierungsqualität führte. Bei einer Batchgröße von eins war die Edge-TPU etwa 6x schneller als eine moderne GPU. Insgesamt haben wir festgestellt, dass Mixed Precision ein geeignetes Tool ist, um durch einfache technische Änderungen die Trainingszeit zu verkürzen und den Energieverbrauch zu senken. So kann die vorhandene Hardware effizienter genutzt und der CO2-Ausstoß gesenkt werden.

Marion Dörrich

References

1 P. Gómez, A. M. Kist, P. Schlegel, D. A. Berry, D. K. Chhetri, S. Dürr, M. Echternach, A. M. Johnson, S. Kniesburges, M. Kunduk et al., "BAGLS, a multihospital benchmark for automatic glottis segmentation," Scientific data, vol. 7, no. 1, p. 186, 2020.

2 O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, 2015, pp. 234–241.

preCICE Workshop 2023 - back to Garching!

After two years online, the preCICE Workshop returned to Garching, organized by the TUM SCCS, and hosted at the very hospitable facilities of the LRZ. As always, the talks have been recorded and will gradually be made available online. But these tell only half the story!

This year, 45 participants attended the workshop between Feb. 13-16, two of which crossing the Atlantic to meet the preCICE community, and many others taking long train rides from all over Europe. The unique combination of hands-on training, developer and user talks, World Café, and user support sessions makes the preCICE Workshop a no-miss.

Figure 1: Group photo of the preCICE Workshop 2023

A nice suprise this year was the broad range of new applications covered by user talks. Yannic Fischler and Daniel Abele from the AWI group of Prof. Angelika Humbert talked about using preCICE to couple ice-sheet models. Prof. Ahmed Elseikh from the Heriot Watt University (UK) demonstrated how they use preCICE to couple OpenFOAM to OpenAI. Leonard Willeke from the Univ. Stuttgart demonstrated coupling with the functional mock-up

interface, opening a large new box of potential projects. Carme Homs Pons (also Univ. Stuttgart) talked about coupling muscle models. Finally, and next to more user and developer talks, the invited speaker Prof. Philip Cardiff from the University College Dublin (Ireland) showed how they are coupling solids4foam with OpenFOAM via preCICE.

Appart from the talks, the community had the opportunity to discuss the present and future of preCICE, giving us important feedback for the upcoming preCICE v3 and our long-term research, as well as to get feedback from the experts for their own projects.

Figure 2: preCICE World Café and dinner on February 14

While it is not yet clear when the next preCICE workshop will be, one thing is clear: there will be one, and it will be even better! See you there.
Read more on `https://precice.discourse.group/t/1397`

Gerasimos Chourdakis

Zwei Jahre NHR@FAU in Erlangen

Wenn man was geschafft hat, darf man auch feiern:
Am **15. März 2023** konnte das Zentrum für Nationales Hochleistungsrechnen Erlangen (NHR@FAU) auf etwas mehr als zwei Jahre Betrieb zurückblicken. Als eines von insgesamt neun nationalen Zentren für das nationale Hochleistungsrechnen (NHR) wurde das NHR@FAU am 1. Januar 2021 aus dem Regionalen Rechenzentrum Erlangen (RRZE) heraus gegründet und stellt über die Friedrich-Alexander-Universität Erlangen-Nürnberg hinaus HPC-Ressourcen und Beratungskompetenz für die gesamte Region und deutschlandweit den Forschenden zur Verfügung.
Zugleich leitet es die gemeinsamen Performance-Engineering-Aktivitäten innerhalb der NHR-Allianz und unterstützt Wissenschaftlerinnen und Wissenschaftler mit spezieller Expertise auf dem Gebiet der atomistischen Simulationen.

Das Event gewährte den Besucherinnen und Besuchern spannende Einblicke in zukunftsweisende interdisziplinäre Forschungsfelder und performante Hochleistungsrechensysteme. In seinem Grußwort betonte Bayerns Innenminister Joachim Herrmann die Wichtigkeit des neuen Zentrums für die wissenschaftliche Wettbewerbsfähigkeit Bayerns und Deutschlands. Auch für die Zukunft ist vorgesorgt: Einem Neubau für ein neues Nordbayerisches Hochleistungsrechenzentrum ist bereits der Weg geebnet.
Mehr dazu ab Seite 26.

Abbildung 1: Prof. Dr. Dieter Kranzlmüller (Vorsitzender des Direktorums des LRZ Garching), Prof. Dr. Gerhard Wellein (Direktor NHR@FAU) und Staatsminister Joachim Herrmann beim Festkolloquium.

Dass sich die Mühen in den vergangenen zwei Jahren gelohnt haben, wurde am 16. März im „NHR@FAU Results Symposium“ demonstriert: Über zwanzig Wissenschaftlerinnen und Wissenschaftler präsentierten die neuen Erkenntnisse, die sie mit den beiden Supercomputern „Fritz“ (ein reines CPU-System) und „Alex“ (GPU-System optimiert für atomistische Simulationen und maschinelles Lernen) seit der Inbetriebnahme im Letzten Jahr erzielen konnten. Weitere Informationen gibt es unter go-nhr.de/Colloquium23.

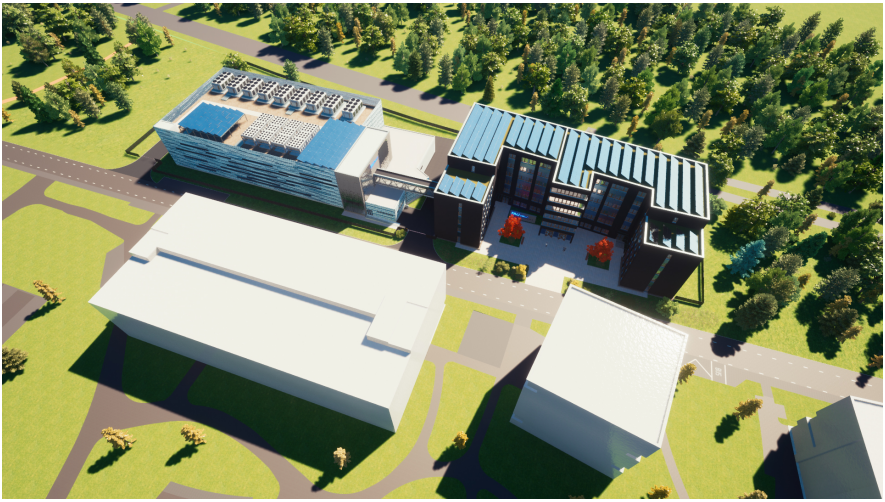
Georg Hager

Breaking Franken News: „Nordbayerische Schwester“ für das Leibniz-Rechenzentrum an der FAU

Mit geballter Ministeriumskraft wurde am 31.5.2023 ein „Jahrhundertprojekt“ für die FAU (Wissenschaftsminister Blume), das über den „Wissenschaftsraum Nürnberg-Erlangen-Fürth hinaus in nationale Versorgungs- und Forschungsstrukturen eingebunden sein wird“ (Bauminister Bernreiter) angekündigt. Innenminister Herrmann verweist darauf, dass sich die „Anziehungskraft der Metropolregion als exzellenter Standort für Wissenschaftler aus der ganzen Welt“ weiter erhöhen wird. Und selbst Finanz- und Heimatminister Füracker sieht einen weiteren wichtigen „Schritt in Richtung digitale Zukunft!“ Ministeriumsübergreifend ist man sich auch einig, dass an der FAU bis Ende des Jahrzehnts die „nordbayerische Schwester“ (O-Ton der Pressemitteilung [1] des Leibniz-Rechenzentrums entstehen soll.

Aber was ist eigentlich passiert? Wissenschaftsminister Markus Blume hat die Freigabe für den Planungsauftrag zum Bau eines nordbayerischen Hochleistungsrechenzentrums erteilt. Bis zu 260 Millionen Euro sollen für dieses Hightech-Vorhaben bereitgestellt werden. Damit erhält nicht nur das NHR@FAU langfristig eine moderne und energieeffiziente Infrastruktur zum Betrieb leistungsfähiger Systeme für HPC- und KI-Anwendungen. Gleichzeitig soll auch ein IT-Leuchtturm für alle Universitäten und Hochschulen in Nordbayern geschaffen werden. Auch bei dieser Perspektive steht der südbayerische Bruder in Garching Pate. Gemeinsam mit den Geschwistern am LRZ haben sich Mitarbeiter des NHR@FAU und des RRZE aktiv am bisherigen Planungsprozess beteiligt und gerade für den geplanten Rechnerraum ein zukunftsweisendes Konzept entworfen, das nun dem Planungsauftrag und dessen Umsetzung zu Grunde liegt.

[1] https://www.stmwk.bayern.de/pressemitteilung/12663/nr-48-vom-31-05-2023.html



Konzeptstudie für Rechnerraum (oben links) und Bürogebäude (oben rechts)

Für das HPC an der FAU und in ganz Bayern ist dies sicherlich ein Meilenstein und wir freuen uns in Franken darauf, dieses Projekt nun konkret planen und dann hoffentlich auch bis zum Ende des Jahrzehnts umsetzen zu können. Ob es dann final eine „nordbayerische Schwester“ oder ein „fränkischer Bruder“ wird, wird sich bis dahin noch entscheiden. Gleiches gilt für die Abgrenzung zwischen Franken und Nordbayern – was ja dann zentral für die „Nutzungsberechtigung“ sein wird. Ich freue mich schon auf die „Gender Reveal Party“!

Trotzdem bedanke ich mich schon jetzt bei den vielen Kollegen (da haben wir leider noch immer ein Gendergap) in Nord- und Südbayern, die das Projekt auf technischer Ebene vorangebracht haben sowie beim Staatsministerium für Wissenschaft und Kunst für die Unterstützung bei dieser finanziellen Herkulesaufgabe.

Gerhard Wellein

BGCE Opening Weekend
A new season

With the arrival of spring comes the Bavarian Graduate School of Computational Engineering (BGCE) program's Opening Weekend (OWE), which took place on the 14th to 16th of April this year. Unlike the previous occasion, this OWE was heralded by clear skies and bright weather; a fortuitous start to the yearly gathering of the BGCE family, professors, seniors and coordinators. This year, we welcome an intake of 14 new juniors that started their studies in the Winter Semester 2022/23.

After some "hard facts" by our very own Dr. Tobias Neckel, the BGCE newcomers broke the ice on Friday, lead by the two coordinators Jan Seydel and Michael Wiedenmann in the Soft Skill Seminar "When teamwork works".

Figure 1: When teamwork works. Photo by Michael Wiedenmann

We enjoyed an in-depth exploration of high-performance circuit simulation with a talk from Dr. Christoph Kowitz representing Infineon in the traditional

“Kaminabend”. On Saturday, the juniors particpated in a “contracting” session, negotiating the expectations, obligations. The seniors took part in the Soft Skill seminar “Step Out”, that focused on team-building.

Figure 2: Signed agreement after contracting. Photo by Jan Seydel

The final day of the Opening Weekend took place on Sunday the 16th of April. There, the juniors and seniors worked through the “Group Challenge” activity together, where they were presented with a complicated task to complete together in a minimal stint of time. This finally culminated in the final “Consulting Circle”, where the seniors could share their gathered wisdom to the juniors in a guided, informal session.

Figure 3: Seniors Step Out. Photo by Stefanie Rein

As has so often been the case in recent years, we fully enjoyed the interaction with this group of motivated and committed students. We are very much looking forward to working with them for another eventful BGCE year.

Keefe Huang

* Notiz * Notiz * Notiz *

Termine 2023

- **Upcoming SIAM Conferences & Deadline**
  https://www.siam.org/conferences/calendar
- **preCICE Minisymposium Greece, Crete, Chania**
  ECCOMAS Coupled Problems 2023: 05.06.23 - 07.06.2023
  https://precice.org/eccomas-coupled-2023.htmlorhttps://coupled2023.cimne.com/
- **Supercomputing 2023:**
  The International Conference for High Performance Computing, Networking, Storage and Analysis (SC23) – SC 23 in Denver, CO, USA: 12.11.-17.11.2023 https://sc23.supercomputing.org/
- **KONWIHR Workshop**
  October 11, 2023 (online) https://www.konwihr.de/

Quartl* - Impressum
Herausgeber:
Prof. Dr. A. Bode, Prof. Dr. H.-J. Bungartz, Prof. Dr. U. Rüde
Redaktion:
S. Herrmann, S. Reiz, Dr. S. Zimmer
Technische Universität München
School of Computation, Information and Technology
Boltzmannstr. 3, 85748 Garching b. München
Tel./Fax: ++49-89-289 18611 / 18607
e-mail: herrmasa@in.tum.de,
https://www.cs.cit.tum.de/sccs/startseite/
Redaktionsschluss für die nächste Ausgabe: **01.09.2023**

* Quartel: früheres bayerisches Flüssigkeitsmaß,
→ das Quart: 1/4 Kanne = 0.27 l
(Brockhaus Enzyklopädie 1972)